UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA, ECOLOGIA E EVOLUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Identificação *de novo* e citogenômica comparativa de DNAs satélites em espécies de *Drosophila* dos grupos *virilis* e *montium*

Orientado: Bráulio Soares Macedo Leão e Silva
Orientador: Dr. Gustavo Campos e Silva Kuhn
Coorientadora: Dra. Marta Svartman

Belo Horizonte
2021

Bráulio Soares Macedo Leão e Silva

Identificação *de novo* e citogenômica comparativa de DNAs satélites em espécies de *Drosophila* dos grupos *virilis* e *montium*

Dissertação apresentada ao programa de Pós-Graduação em Genética da Universidade Federal de Minas Gerais como pré-requisito obrigatório para obtenção do título de Mestre em Genética, área de concentração Genômica e Bioinformática.

Orientador: Dr. Gustavo C.S. Kuhn
Coorientadora: Dra. Marta Svartman

Belo Horizonte

2021

043 Leão e Silva, Bráulio Soares Macedo.
Identificação de novo e citogenômica comparativa de DNAs satélites em espécies de Drosophila dos grupos virilis e montium [manuscrito] / Bráulio Soares Macedo Leão e Silva. - 2021.
105 f. : il. ; 29,5 cm.

Orientador: Dr. Gustavo Campos e Silva Kuhn. Coorientadora: Dra. Marta Svartman.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Citogenética. 3. Drosophila. 4. DNA Satélite. 5. Elementos de DNA Transponíveis. 6. Biologia Computacional. I. Kuhn, Gustavo Campos e Silva. II. Svartman, Marta. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título

CDU: 575

Ficha catalográfica elaborada por Fabiane C. M. Reis - CRB: 6/2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

ATA DE DEFESA DE DISSERTAÇÃO / TESE

ATA DA DEFESA DE DISSERTAÇÃO	**313/2021** **entrada**
Bráulio Soares Macedo Leão e Silva	**1º/2019** **CPF: 125.107.876-16**

Às quatorze horas do dia **20 de agosto de 2021**, reuniu-se remotamente, devido ao isolamento social decorrente da pandemia de COVID-19, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Identificação de novo e citogenômica comparativa de DNAs satélites em espécies de Drosophila dos grupos virilis e montium"**, requisito para obtenção do grau de Mestre em **Genética.** Abrindo a sessão, o Presidente da Comissão, **Gustavo Campos e Silva Kuhn**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**CPF**	**Indicação**
Gustavo Campos e Silva Kuhn	UFMG	260.136.648-62	Aprovado
Marta Svartman	UFMG	101.787.258-97	Aprovado
Francisco Pereira Lobo	UFMG	012.273.736-94	Aprovado
Mateus Mondin	ESALQ/USP	264.195.108-80	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado.**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 20 de agosto de 2021.

Gustavo Campos e Silva Kuhn (UFMG)

Marta Svartman (UFMG)

Francisco Pereira Lobo (UFMG)

Mateus Mondin (ESALQ/USP)

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Mateus Mondin**, **Usuário Externo**, em 20/08/2021, às 17:34, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn**, **Professor do Magistério Superior**, em 20/08/2021, às 17:36, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 20/08/2021, às 17:38, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marta Svartman**, **Professora do Magistério Superior**, em 23/08/2021, às 08:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0911146** e o código CRC **C35E107D**.

Referência: Processo nº 23072.243468/2021-81 SEI nº 0911146

Dedico este trabalho aos animais sacrificados em prol da ciência.

Agradecimentos

Agradeço ao meu orientador Prof. Dr. Gustavo Campos e Silva Kuhn e à minha coorientadora Profa. Dra. Marta Svartman por todos os ensinamentos, conselhos e momentos vividos durante esses 5 anos em que estive no laboratório. Hoje, concluo mais uma etapa da minha vida profissional e pessoal que, sem a presença de vocês, não seria possível. A pesquisa científica é um livro sem fim, e, ter profissionais tão competentes ao meu lado, só engrandece o brilho deste capítulo que escrevo.

Às bancas examinadoras da dissertação e da disciplina de seminários, por terem aceitado ler e discutir o meu trabalho.

À minha família, em especial, minha mãe Ana Maria, meu pai Sérgio Ricardo e meu irmão Samuel, por sempre terem me apoiado nessa jornada e por acreditarem no meu potencial. Sou uma pessoa iluminada pois tenho vocês comigo!

À minha namorada Carolina, que compartilha alegrias e tristezas comigo há mais de 9 anos. Obrigado por estar ao meu lado e me mostrar, dia após dia, que a vida é muito mais bonita e prazerosa com você.

Aos meus queridos amigos do laboratório, Rafa, Erick, Ana, Pedro, Zé, Gui, Mirela, Radarane, Alice e Naiara. Com toda certeza, o LCEv foi o melhor lugar que já trabalhei, e isso graças a vocês, que fizeram meu cotidiano menos difícil e mais descontraído. Ter conhecido e trabalhado com vocês foi uma das maiores conquistas que tive nos meus anos de ICB.

Aos meus professores e todos os funcionários da Pós-graduação em Genética. Especialmente, ao Daniel, que foi essencial para o desenvolvimento do meu projeto.

Às amizades que construí na graduação: Matheus, Camila, Pedro e Rodrigo. Mesmo que o tempo e a distância não nos mantenham mais juntos diariamente, vocês sempre foram importantes na minha caminhada. Sou grato pela amizade e carinho compartilhados!

Aos amigos da Pós, principalmente os integrantes do “Trem Bão é Ciência”. Nosso projeto é muito bonito e tem crescido cada vez mais. Obrigado por não me deixarem esquecer que a ciência feita nos laboratórios e institutos de pesquisa brasileiros precisa ser levada e explicada à população, sua verdadeira fonte de investimento.

À CAPES, FAPEMIG e CNPq, agências financiadoras do meu e de outros projetos do laboratório. À CAPES, sobretudo pela bolsa de pesquisa concedida e

aceite de prorrogação do prazo de defesa devido à pandemia da COVID.

É difícil lembrar de tudo e de todos que, de alguma maneira, fizeram parte deste processo. Mas, para você leitor, meu muito obrigado por despender seu precioso tempo lendo esta dissertação. A você, obrigado por acreditar na ciência em tempos sombrios.

Sumário

Lista de figuras

Lista de abreviaturas

BLAST – “Basic Local Alignment Search Tool” (Ferramenta de busca e alinhamento básico local)
De novo – novo, do começo
DNA - ácido desoxirribonucleico (ADN)
e.g. – “Exempli gratia” (por exemplo)
FISH – Fluorescence in situ hybridization (Hibridização *in situ* fluorescente)
HKY – Hasegawa-Kishino-Yano model
i.e. – “Id est” (isto é)
In silico – em ou de ambiente computacional
In situ – no seu lugar de origem, local de origem
JC – Jukes-Cantor model
K2 – Kimura 2-parameter model
Kbs – kilobases
MEGA - Molecular Evolutionary Genetics Analysis (software)
NCBI – National Center for Biotechnology Information (instituição)
NGS – “Next-generation sequencing” (sequenciamento de nova geração)
Pb/bp – pares de base/”base pairs”
PCR – reação em cadeia da polimerase
Reads – fragmentos gerados após sequenciamento de DNA
satDNAs – DNAs satélites
T92 – Tamura 3-parameter model
TEs – Elementos transponíveis
R – Linguagem de programação

Resumo

Os DNAs repetitivos são os componentes mais abundantes dos genomas de eucariotos, podendo ser classificados como dispersos, como é o caso dos elementos transponíveis (TEs), ou em tandem, como no caso dos microssatélites, minissatélites e DNAs satélites (satDNAs). Os estudos de DNAs repetitivos durante a era pré-genômica eram muitas vezes demorados, trabalhosos e limitados. O desenvolvimento de técnicas de sequenciamento massivo de DNA e a disponibilidade de várias ferramentas de bioinformática propiciam atualmente o acesso e anotação automatizada de praticamente todos os DNAs repetitivos presentes em um genoma. Desta forma, torna-se necessário testar a capacidade destas ferramentas para a correta identificação e classificação destas sequências repetitivas. No presente projeto, utilizamos os pipelines RepeatExplorer e TAREAN para a identificação *de novo* de satDNAs em duas espécies próximas de *Drosophila* do grupo *virilis* (*D. virilis* e *D. americana*) e em 23 espécies recém sequenciadas do grupo *montium*. Enquanto *D. virilis* e *D. americana* já foram muito estudadas no contexto de DNAs repetitivos na era pré-genômica, nenhuma espécie do grupo *montium* havia sido investigada nesse contexto. Em nossas análises, identificamos seis famílias de elementos repetitivos em tandem (TRs) com características de satDNAs nas espécies do grupo *virilis* e 142 clusters de satDNAs nas espécies do grupo *montium*. Para *D. virilis* e *D. americana*, estudamos em detalhe cada uma destas famílias, combinando dados da literatura, dados gerados pelos *pipelines* testados e dados novos sobre localização cromossômica. Para o grupo *montium*, realizamos a identificação *de novo* com o TAREAN e uma investigação mais detalhada das famílias de satDNAs compartilhadas entre as espécies. Conseguimos identificar e caracterizar os “reais” satDNAs de *D. virilis* e *D. americana*, além de analisar relações evolutivas entre satDNAs e elementos transponíveis. Para o grupo *montium*, fizemos pela primeira vez uma identificação de sequências de satDNAs, demonstramos que alguns satélites compartilham sequências com elementos transponíveis, e que famílias de satDNAs podem ser utilizadas como marcadores taxonômicos e filogenéticos dentro do grupo. Sendo assim, confirmamos a eficácia dos pipelines RepeatExplorer e TAREAN na identificação *de novo* de satDNAs em *Drosophila*, mostramos a importância do uso combinado de abordagens *in silico* e experimentais para a identificação e caracterização de satDNAs, relatamos exemplos, dentro do grupo *montium*, onde satDNAs podem ser úteis como marcadores filogenéticos e, por último, relatamos vários tipos de associações interessantes entre satDNAs e TEs.

Palavras-chave: DNAs repetitivos, DNAs satélites, Elementos transponíveis, citogenômica, *Drosophila*, RepeatExplorer, TAREAN, grupo *virilis*, grupo *montium*

Abstract

Repetitive DNAs are the most abundant components of the eukaryotic genomes and can be classified as dispersed, as in the case of transposable elements (TEs), or in tandem, as in the case of microsatellites, minisatellites and satellite DNAs (satDNAs). During the pre-genomic era, repetitive DNA studies were often time-consuming, laborious and limited. Currently, the development of massive DNA sequencing techniques and the availability of several bioinformatics tools provide access and automated annotation of virtually all repetitive DNAs present in a genome. Thus, it is necessary to test the capacity of these tools for the correct identification and classification of these repetitive sequences. In the present project, we used the RepeatExplorer and TAREAN pipelines for *de novo* identification of satDNAs in two closely species of *Drosophila* from the *virilis* group (*D. virilis* and *D. americana*) and in 23 newly sequenced species from the *montium* group. While *D. virilis* and *D. americana* have been extensively studied in the context of repetitive DNA in the pre-genomic era, no species from the *montium* group had been previously investigated in this context. In our analyzes, we identified six families of repetitive tandem repeats (TRs) with satDNA features in the *virilis* group species and 142 satDNA clusters in the *montium* group species. For *D. virilis* and *D. americana*, we studied each of these families in detail, combining data from the literature, data generated by the tested pipelines and new data on chromosomal localization. For the *montium* group, we carried out the identification with TAREAN and a more detailed investigation of the satDNA families shared between species. We were able to identify and characterize the “real” satDNAs of *D. virilis* and *D. americana*, in addition to analyze evolutionary relationships between satDNAs and transposable elements. For the *montium* group, we did an unprecedented identification of satDNA sequences, demonstrating that some satellites share sequences with transposable elements and that satDNA families can be used as taxonomic and phylogenetic markers within the group. Thus, we confirmed the effectiveness of the RepeatExplorer and TAREAN pipelines for *de novo* identification of satDNAs in *Drosophila*, we showed the importance of combining *in silico* and experimental approaches for the identification and characterization of satDNAs, we reported examples, within the *montium* group, where satDNAs may be useful as phylogenetic markers and lastly, we reported several types of interesting associations between satDNAs and TEs.

Keywords: repetitive DNAs, satellite DNAs, Transposable elements, cytogenomics, *Drosophila*, RepeatExplorer, TAREAN, *virilis* group, *montium* group

1. Introdução

1.1. DNAs satélites: elementos repetitivos abundantes dos genomas eucariotos

Os genomas nucleares de organismos eucariotos são complexos e apresentam diferentes tipos de sequências de DNA. Por alguns anos, no século passado, pensava-se que esses genomas eram formados principalmente por sequências codificadoras funcionais, como por exemplo, os genes. Entretanto, com a evolução de estudos na área da genética e, principalmente, após o desenvolvimento de técnicas de sequenciamento de DNA, descobriu-se que, na verdade, grande parte dos genomas de eucariotos é enriquecida com sequências de DNA repetitivos não-codificadores (Britten e Kohne, 1968; Charlesworth e col. 1994; Gregory, 2005).

De acordo com sua organização, as sequências repetitivas podem ser classificadas em dispersas ou em tandem (i.e., uma cópia seguida da outra). Como exemplo de sequências dispersas, pode-se citar os elementos transponíveis (TEs), que são elementos abundantes com capacidade de movimentação dentro do genoma. Os TEs podem ser classificados como retrotransposons (TEs de classe I) ou como transposons (TEs de classe II) (Charlesworth e col. 1994). Essa classificação é baseada nos mecanismos de transposição dessas sequências, sendo eles: transposição via RNA intermediário, realizada pelos retrotransposons; e transposição direta via sequência de DNA, utilizada pelos transposons de DNA. Além disso, esses elementos podem ser classificados como autônomos, quando codificam as enzimas mediadoras da transposição, e não-autônomos, quando utilizam as enzimas de outros TEs para mobilização (Griffiths e col. 2016).

Entre as repetições organizadas em tandem estão os RNAs ribossomais, algumas famílias de genes codificadores de proteínas, microsatélites, minissatélites e DNAs satélites (satDNAs) (López-Flores e Garrido-Ramos, 2012; Biscotti e col. 2015). Os microsatélites, minissatélites e DNAs satélites se diferenciam principalmente pelo seu tamanho, número de repetições e localização das cópias nos genomas. De acordo com Tautz (1993), os microsatélites ou sequências simples possuem monômeros (i.e. unidades de repetição) entre 1-6 pb, que podem ser encontrados em qualquer parte do genoma e se repetem de 5-100x em cada arranjo. Já os minisatélites podem variar entre 9-100 pb, estão presentes em diferentes regiões nos genomas (mais encontrados nas regiões teloméricas), e se repetem centenas de vezes em cada locus. Os satDNAs, entretanto, são geralmente encontrados na heterocromatina, principalmente nas regiões centroméricas, onde estão organizados em arranjos que

podem alcançar vários kbs (kilobases) (Figura 1). Esses elementos possuem tamanho variado, com monômeros que podem ter entre 2 pb a mais de 1.000 pb (>1kb) (Miklos, 1985; Plohl e col. 2012).

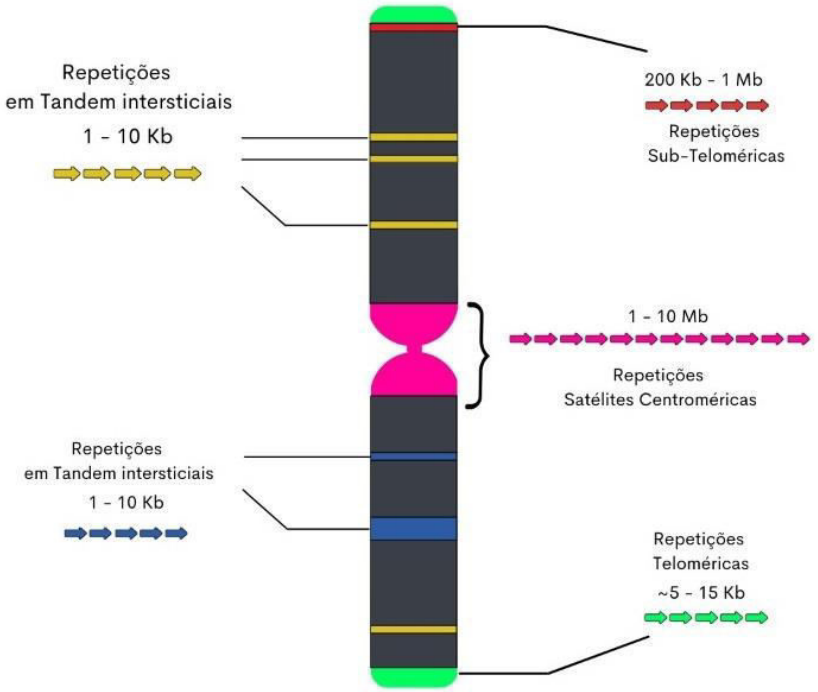


Figura 1. Cromossomos de organismos eucariotos são enriquecidos com repetições de elementos repetitivos em tandem, como os DNAs satélites. Grandes arranjos de DNAs satélites são encontrados em regiões de heterocromatina, como por exemplo no centrômero e regiões sub-teloméricas. Além disso, sequências satélites também podem ser encontradas em arranjos pequenos dispersos ao longo dos braços eucromáticos (Ugarković e Plohl, 2002; Heslop-Harrison e col. 2003; Mravinac e col. 2004; Kuhn e col. 2008, 2012).

O número de famílias de satDNAs varia bastante entre os organismos: enquanto algumas espécies apresentam maior diversidade, como o gafanhoto *Locusta migratória* que possui 62 famílias, outras possuem um número bem menor, como por exemplo o *Homo sapiens*, que possui 9 famílias de satDNAs identificadas (Ruiz-Ruano e col. 2016; Garrido-Ramos, 2017). No caso do genoma humano, por exemplo, há um predomínio da família satDNA α, que representa mais da metade dos satDNAs encontrados no nosso genoma. Já em espécies de plantas, uma mesma família de satDNA pode representar entre 0,1% e 36% dos genomas, como no caso da abundante família FriSAT1, presente no gênero *Fritillaria* (Ambrožová e col. 2011).

DNAs satélites possuem uma taxa rápida de evolução molecular, sendo frequentemente compartilhados entre espécies próximas filogeneticamente. Uma vez que essas sequências não possuem função codificadora, é esperado que estejam evoluindo de forma neutra, ou seja, que ao longo do tempo as cópias se diversifiquem,

criando novas variantes. Entretanto, esta não é a situação observada. Ao contrário, cópias de DNAs satélites presentes em uma mesma espécie são muito similares, indicando um processo de homogeneização das sequências. Este padrão interessante de evolução molecular é bastante comum em sequências repetidas em tandem e se tornou conhecido como evolução combinada (ou evolução em concerto) (Dover, 1982).

A evolução combinada (Figura 2) é um processo contínuo e ocorre entre cópias presentes em diferentes partes do genoma, sendo consequência, principalmente, de mecanismos de conversão gênica, *crossing-over* desigual, transposição, derrapagem da polimerase durante a replicação do DNA e trocas mediadas por RNA (Smith, 1976; Zimmer e col. 1980; Dover, 1982, 1994; Thompson-Stewart e col. 1994; Cohen e Segal, 2009).

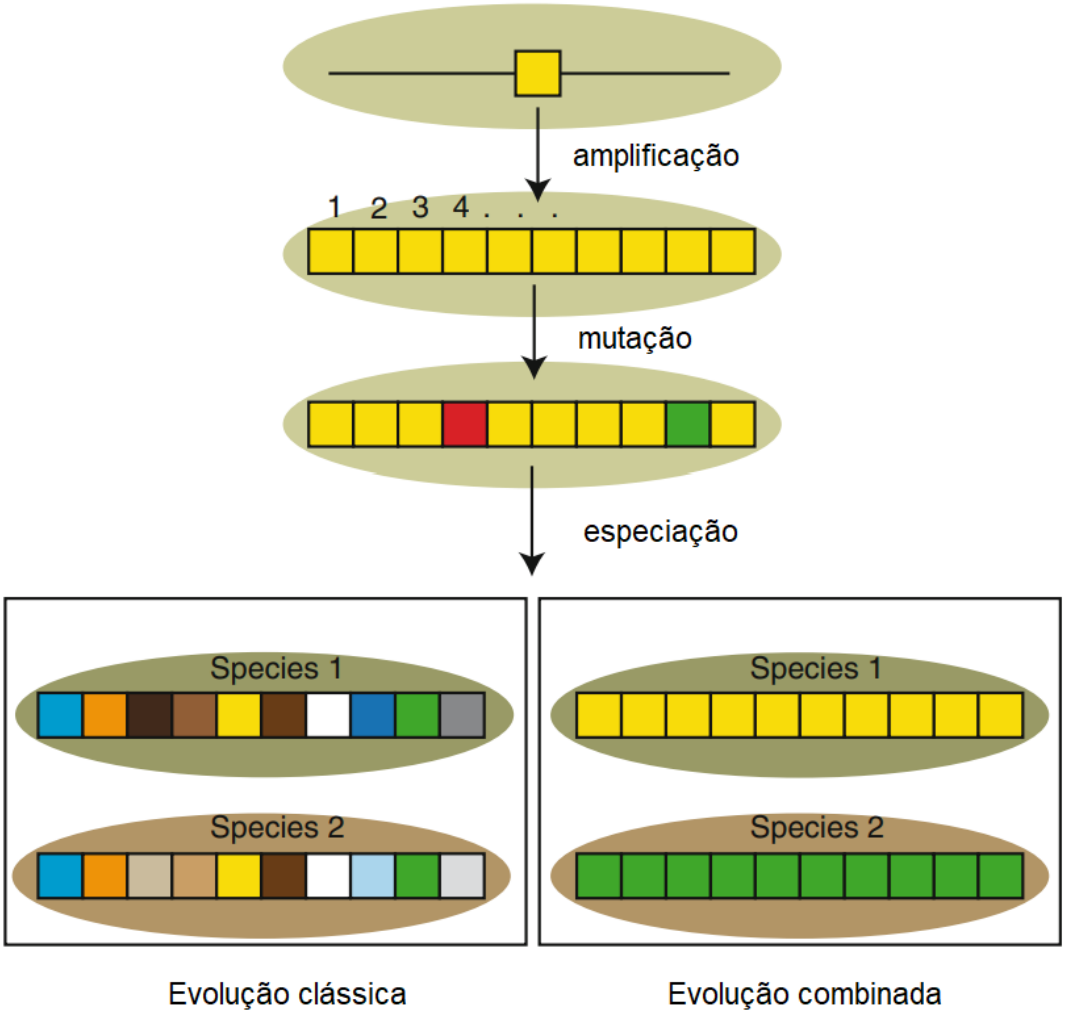


Figura 2. Modelo representativo do processo de evolução combinada para sequências em tandem (direita) (Adaptado de Feliner e Rossello, 2012). Mutações geram variação entre as cópias. Algumas destas mutações podem ser homogeneizadas nos arranjos, através da ação

de mecanismos de recombinação desigual. Mutações diferentes podem ser homogeneizadas e fixas em diferentes espécies.

Uma mesma família de satDNAs pode apresentar mais de uma subfamília, ou seja, arranjos contendo diferentes variantes homogeneizadas da mesma família. Essas subfamílias podem estar presentes em um mesmo genoma, ou em genomas de espécies diferentes (Palomeque e Lorite, 2008; Garrido-Ramos, 2017; Louzada e col. 2020).

Os DNAs satélites não codificam proteínas, no entanto, podem exercer papéis importantes no genoma, como no processamento de RNAs, na regulação de compensação de dose no cromossomo X em machos, na modulação da cromatina e na formação do centrômero e do cinetócoro (Henikoff e col. 2001; Volpe e col. 2002; Ugarković, 2005; Menon e col. 2014; Rošić e col. 2014) (Figura 3).

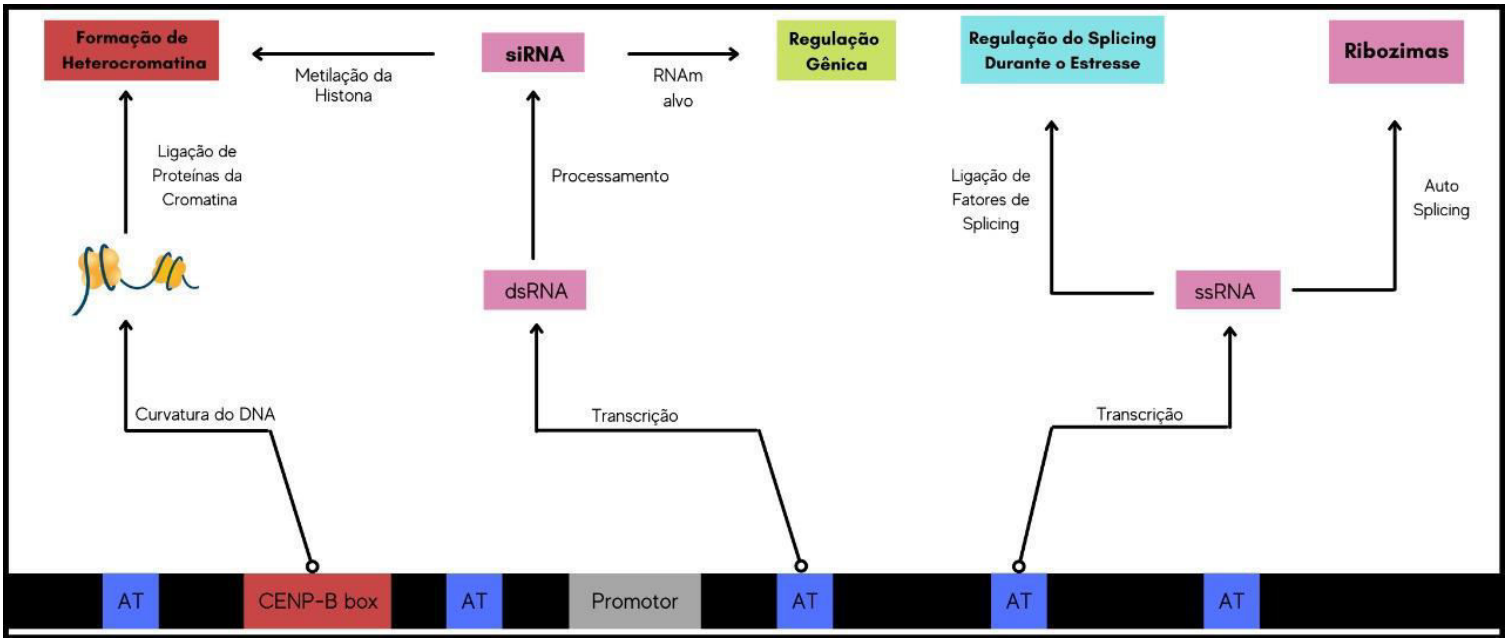


Figura 3. Processos biológicos em que DNAs satélites podem estar envolvidos em células eucarióticas. (Figura adaptada de Ugarković, 2005).

Os primeiros estudos com DNAs satélites tiveram início há mais de 50 anos atrás, quando o isolamento das sequências repetitivas era realizado por meio de experimentos de ultracentrifugação em gradiente de densidade contendo cloreto de césio (Kit, 1961; Gall e col. 1971). Entretanto, técnicas como essa eram muito limitadas, uma vez que se restringiam a identificações superficiais de satDNAs. Ao longo dos anos, novas tecnologias foram desenvolvidas e incorporadas a estudos de satDNAs, como por exemplo, a digestão de satélites por enzimas de restrição (endonucleases) que realizavam a fragmentação/isolamento dessas sequências.

Além desses métodos, análises com *dot-blot*, *Southern Blot* e hibridização *in situ* (como a *Fluorescence in situ hybridization* - FISH) se popularizaram e vem sendo utilizadas ao longo dos anos para a identificação e caracterização mais precisa de satDNAs de diferentes genomas (Castagnone-Sereno e col. 2008; Ruiz-Ruano e col. 2016; Utsunomia e col. 2017; Gatto e col. 2018). A FISH com DNAs satélites, por exemplo, é uma técnica de citogenética molecular utilizada para a localização de satDNAs em cromossomos. Consiste em um método onde o satDNA é isolado, marcado com molécula fluorescente e, posteriormente, hibridado por complementaridade de sequência em cromossomos metafásicos, politênicos, núcleos interfásicos e/ou fibras de DNA estendidas preparadas em lâmina de microscopia. Após a hibridação, é possível analisar a distribuição cromossômica do satDNA alvo através de microscopia de fluorescência.

Entretanto, mesmo com a disponibilidade de técnicas de citogenética e biologia molecular para estudos de satDNAs, esses métodos são enviesados, o que não permite a detecção completa do conjunto de satDNAs (satelitoma) do genoma e, além disso, se restringem a análises de poucas sequências para cada satélite descoberto. Esta lacuna começou a ser preenchida após o desenvolvimento de técnicas de sequenciamento de DNA em massa, em especial, das tecnologias de *Next-Generation Sequencing* (NGS) (Garrido-Ramos, 2017; Lower e col. 2018).

A era do "*big data*" revolucionou os estudos dentro da área da genética, e atualmente, já está bem difundida ao redor do mundo e em diversas áreas de pesquisa. Inicialmente, apenas genomas pequenos eram sequenciados, porém, ao longo dos anos, as tecnologias começaram a suportar maior quantidade de dados, também analisados em menor tempo (van Dijk e col. 2014). Neste contexto, destaca-se o Projeto Genoma Humano (PGH), que no fim do século XX foi um esforço científico internacional cujo objetivo principal era realizar o primeiro sequenciamento de um genoma humano (Lander e col. 2001; Venter e col. 2001). Entre outros eucariotos, *Caenorhabditis elegans* foi o primeiro organismo multicelular cujo genoma foi sequenciado (The *C. elegans* Sequencing Consortium, 1998). Além deste nematódeo, destaca-se o sequenciamento de *Drosophila melanogaster*, popularmente conhecida como "mosca das frutas" ou "mosca da banana" (Adams e col. 2000), que é um organismo modelo em várias áreas da ciência, em especial na genética.

1.2. O gênero *Drosophila*

Atualmente, mais de 1.600 espécies de *Drosophila* foram descritas, porém estima-se que o número de espécies existentes seja bem maior (Powell, 1997; O'Grady e DeSalle, 2018). O gênero conta com moscas que variam desde o

tamanho do genoma à coloração e tamanho corporal (Markow e O'Grady, 2005; Boulesteix e col. 2006; Bosco e col. 2007).

A história filogenética da família Drosophilidae, onde se encontra o gênero *Drosophila*, ainda é tema de estudos (Remsen e O'Grady, 2002; Da Lage e col. 2007; Russo e col. 2013; Yassin 2013, 2018; O'Grady e DeSalle, 2018). Atualmente, o gênero é tratado como parafilético, com alguns autores sugerindo a divisão do grupo em um ou mais gêneros. Neste sentido, é necessária uma revisão taxonômica detalhada para resolver os conflitos entre dados de taxonomia tradicional e filogenética molecular do grupo (Yassin, 2013).

Recentemente, um catálogo taxonômico da família Drosophilidae listou 3.962 espécies descritas, sendo elas distribuídas em 70 gêneros (Brake e Bächli, 2008; O'Grady e DeSalle, 2018). Embora a classificação de alguns táxons não seja bem definida, é geralmente aceito que todos os gêneros estão incluídos em duas subfamílias, Steganinae e Drosophilinae (Grimaldi e col. 1990). O gênero *Drosophila* está inserido dentro de Drosophilinae e corresponde a aproximadamente 50% das espécies da subfamília (O'Grady e DeSalle, 2018).

Após o primeiro sequenciamento do genoma *D. melanogaster* no ano 2000, vários projetos colaborativos realizaram o sequenciamento do genoma de outras espécies de *Drosophila* (Richards e col. 2005; Drosophila 12 Genomes Consortium col, 2007; Bronski e col. 2020; Conner e col. 2021). Em 2005, por exemplo, Richards e cols. sequenciaram *D. pseudoobscura* e, dois anos depois, um consórcio de pesquisa global sequenciou mais 12 espécies do gênero (*D. sechellia*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* e *D. grimshawi*) (Drosophila 12 Genomes Consortium col, 2007). Atualmente (em 2021), de acordo com o banco de dados genômicos do *National Center for Biotechnology Information* (NCBI), pelo menos 114 espécies de *Drosophila* tiveram seus genomas sequenciados, sendo que a maioria destes genomas se encontra disponível em bancos de dados públicos.

Sendo assim, a alta disponibilidade de dados genômicos completos de drosofilídeos permite estudos mais detalhados a respeito dos satDNAs presentes nos genomas dessas espécies. Em especial, a identificação, caracterização e organização desses elementos repetitivos pode ser explorada com mais detalhes, já que é possível realizar análises comparativas entre espécies.

1.2.1. Os grupos *virilis* e *montium* do gênero *Drosophila*

A grande variedade de espécies de moscas do gênero *Drosophila* fez com que os "drosofilistas" criassem categorias taxonômicas adicionais para o gênero. Desta

maneira, após as categorias de “gênero” e “subgênero”, há também “grupos”, “complexos” e “clusters” de espécies de *Drosophila*. Estima-se que os subgêneros *Drosophila* e *Sophophora* tenham se separado há aproximadamente 40 milhões de anos (Powell e DeSalle, 1995; Powell, 1997). Após a separação, novos grupos se originaram por irradiação adaptativa, o que levou à origem das mais diversas espécies encontradas atualmente (O’Grady e DeSalle, 2018). Alguns desses grupos são demonstrados na figura 4, com ênfase para espécies do grupo *virilis* e *montium*, foco de nosso estudo.

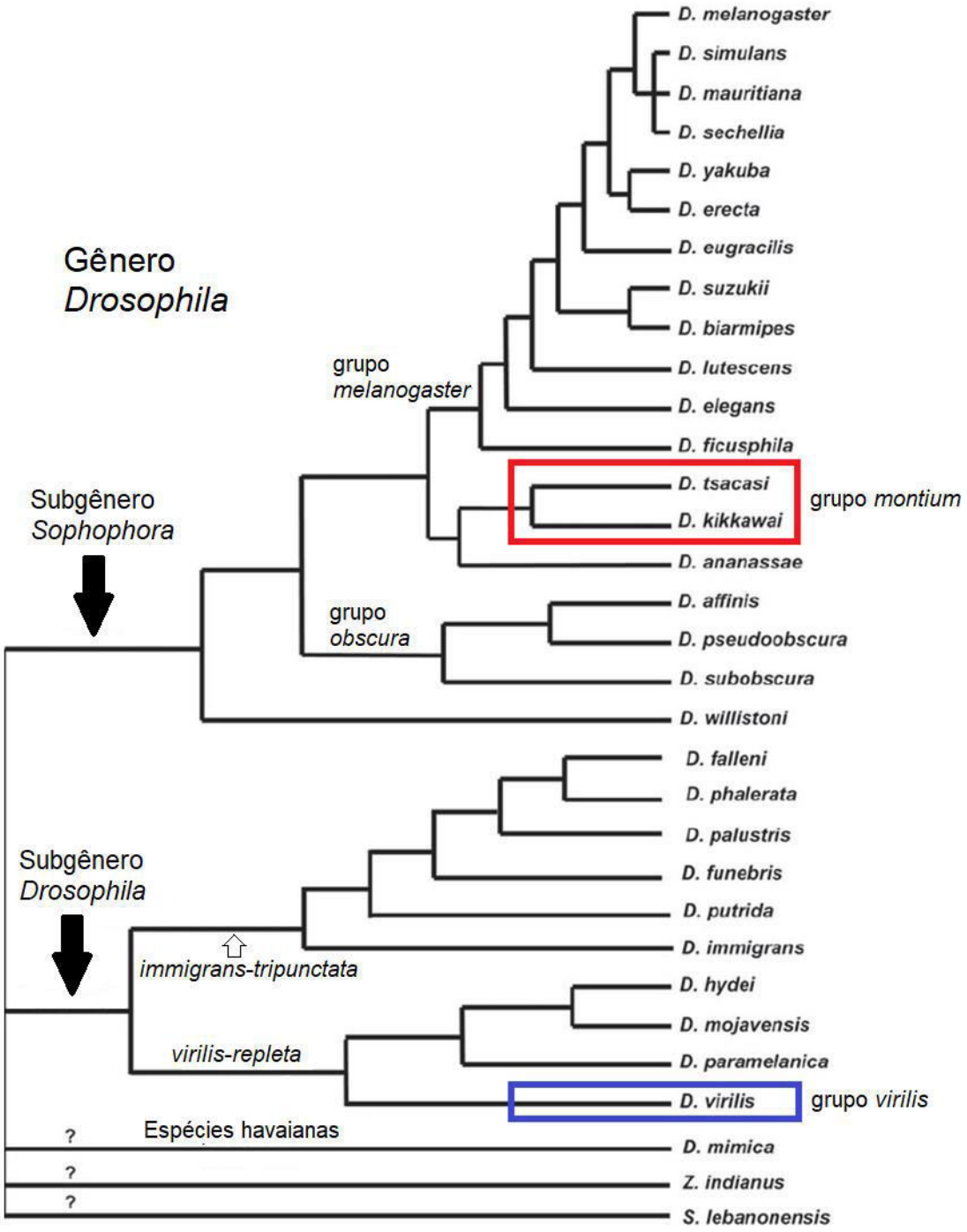


Figura 4. Filogenia demonstrando as relações de parentesco entre os principais grupos do gênero *Drosophila* (Adaptado de Kacsoh e col. 2014). Estão evidenciadas espécies dos grupos

montium (vermelho) e *virilis* (azul). Os grupos *virilis* e *montium* compartilham um ancestral comum há ~40 milhões de anos atrás, sendo considerados dois grupos distantes filogeneticamente dentro do gênero *Drosophila*.

O grupo *virilis* é um dos maiores grupos do subgênero *Drosophila*. Atualmente, compreende 13 espécies que se distribuem em diferentes regiões do globo. Inicialmente, o grupo foi subdividido em dois clados: *virilis* e *montana* (Throckmorton, 1982; Spicer, 1991). Porém, após novos estudos, a taxonomia do grupo foi revista e inclui atualmente quatro linhagens: *virilis* (*D. virilis*, *D. lummei*, *D. americana* e *D. novamexicana*), *montana* (*D. flavomontana*, *D. montana*, *D. lacicola* e *D. borealis*), *kanekoi* (formado apenas por *D. kanekoi*) e *littoralis* (*D. littoralis*, *D. ezoana* e *D. canadiana*). Mesmo assim, as relações filogenéticas entre e dentro das linhagens não estão bem definidas, devido ao baixo suporte, em árvores filogenéticas, dos nós basais (Morales-Hojas e col. 2011).

É no grupo *virilis* onde se encontra *D. virilis*, uma espécie que apresenta um dos maiores genomas de *Drosophila* descritos. Em *Drosophila*, o tamanho dos genomas pode variar entre espécies bem como entre diferentes linhagens de uma mesma espécie, sendo que, estimativas feitas com citometria de fluxo revelaram que essa variação é bastante significativa. Por exemplo, enquanto *D. virilis* apresenta genoma médio de 364 Mb, *D. mercatorum* (subgrupo *repleta*) possui um genoma quase 3x menor, com aproximadamente 128 Mb (Bosco e col. 2007). Essa variação é em parte atribuída aos DNAs satélites, dado que quanto maior o genoma, maior a quantidade de DNAs satélites (Bosco e col. 2007). Além disso, essa correlação positiva está de acordo com Hartl (2000), quem sugeriu que a variação no tamanho dos genomas é um reflexo direto de aneuploidias, eventos de transposição de TEs e contrações/expansões derivadas de ganhos e perdas de DNAs satélites em regiões de heterocromatina.

Atualmente, no banco de dados do NCBI, há vários genomas completos disponíveis de espécies do grupo *virilis*, resultados de sequenciamentos feitos com sequenciadores ABID SOLID, Illumina, Oxford Nanopore e Pacbio SMRT (*Single-molecule real-time sequencing*). Dentre eles, destacam-se os genomas de *D. virilis*, *D. americana*, *D. novamexicana*, *D. littoralis*, *D. lummei* e *D. montana*.

O grupo *montium* faz parte do subgênero *Sophophora*, e assim como o grupo *virilis*, apresenta muitas espécies de *Drosophila*. Atualmente, é composto por 94 espécies de origem asiática e australiana (Yassin, 2018). Durante anos, o grupo foi classificado como um subgrupo do grupo *melanogaster*, mas análises posteriores reposicionaram o clado como um grupo independente (Da Lage e col. 2007). Hoje em dia, o grupo se divide em 8 complexos, mas as relações filogenéticas entre algumas

espécies e clados não são bem definidas ainda. Yassin (2018) analisou traços morfológicos e corológicos e subdividiu o grupo em 7 subgrupos: *parvula*, *montium*, *punjabiensis*, *serrata*, *kikkawai e seguyi*.

Desde a década de 1980, estudos genéticos com espécies do grupo *montium* se basearam, principalmente, em análises comparativas de cariótipos. Baimai (1980), por exemplo, encontrou alta variabilidade cromossômica em algumas espécies analisadas, sendo que as maiores variações ocorriam com os cromossomos 4 (microcromossomo) e Y, enquanto o X se mostrou pouco variável. Neste caso, o autor concluiu que a diversidade cariotípica encontrada estava relacionada a diferentes ganhos de heterocromatina nas espécies, onde satDNAs residem.

Até 2020, poucos genomas de espécies do grupo *montium* haviam sido sequenciados (Chen e col. 2014; Allen e col. 2017; Miller e col. 2018; Kim e col. 2021). Entretanto, recentemente, mais de 40 genomas do grupo foram sequenciados por Bronski e col. (2020) e Conner e col. (2021), por meio de tecnologia Illumina HiSeq 2000 e HiSeq 2500 Systems. Segundo os autores, os genomas sequenciados apresentam alta diversidade com relação ao tamanho, sequências repetitivas e níveis de heterozigosidade. Desta maneira, estudos comparativos a partir dos novos dados gerados são necessários tanto para o debate filogenético do grupo, quanto para investigações a respeito da organização, composição e evolução destes genomas.

1.3. Breve histórico do estudo de DNAs satélites em *Drosophila*

Os primeiros estudos de DNAs satélites realizados em espécies do gênero *Drosophila* foram publicados no final dos anos 1960 e início dos anos 1970 (Laird e McCarthy, 1968; Gall e col. 1971; Gall e Atherton, 1974). Estes trabalhos eram focados, principalmente, na identificação e comparação de densidades nucleotídicas dos DNAs satélites das espécies analisadas. Como relatado anteriormente, os principais métodos investigativos utilizados inicialmente eram a centrifugação por diferença de densidade de cloreto de césio (CsCl) e a hibridação *in situ*. Neste contexto, as primeiras espécies analisadas foram *D. virilis*, *D. melanogaster*, *D. simulans*, *D. funebris* e *D. hydei* (Gall e Atherton, 1974; Barnes e col. 1978; Renkawitz, 1979).

Os trabalhos realizados por Gal e cols. são, até hoje, referência para estudos de DNAs satélites em espécies do grupo *virilis*. Estes trabalhos analisaram uma abundante família de satDNAs com monômeros de apenas 7 pb (heptanucleotídeos), e que apresenta quatro variantes que se diferem por apenas uma substituição nucleotídica, sendo elas: Sat1 (AAACTAC), Sat2 (AAA**T**TAC), Sat3 (AAACTA**T**) e Sat4 (AAAC**A**AC). Além disso, foi estimado que essa família de satDNAs ocupa ~40% do

genoma de *D. virilis* e que as variantes estão presentes nas regiões heterocromáticas próximas aos centrômeros de todos os cromossomos, exceto do cromossomo Y (Gall e col. 1971; Gall e Atherton, 1974). Posteriormente, foi identificada a presença da mesma família em todas as espécies do grupo *virilis*, exceto em *D. montana* e *D. littoralis*. Em algumas espécies, também foi relatada a presença de cópias em regiões eucromáticas (Cohen e Bowman, 1979; Cohen e Kaplan, 1982).

Ainda no grupo *virilis*, outras famílias de satDNAs foram identificadas e caracterizadas posteriormente. Aqui, destaca-se o trabalho de Abdurashitov e col. (2013), que identificaram *in silico* e *in situ* (para algumas famílias) sequências satélites de 225 pb, 154 pb e 172 pb nos genomas de *D. virilis* e *D. americana*. Além dessas sequências, uma família com monômeros de 370 pb, denominada PvB370, foi identificada e caracterizada *in situ* em *D. virilis*, *D. americana*, *D. montana*, *D. novamexicana*, *D. lummei* e *D. littoralis* (Heikkinen e col. 1995; Biessmann e col. 2000). Experimentos de hibridação *in situ* mostraram que PvB370 é uma família abundante nos genomas das espécies do grupo *virilis*, e que as cópias se encontram predominantemente nas regiões subteloméricas dos cromossomos.

Em outros grupos e espécies do gênero, famílias de satDNAs de diferentes tamanhos também já foram descritas, como por exemplo: uma família com sequências monoméricas de 180 pb foi descrita em *Drosophila ambigua*, *D. tristis* e *D. obscura*, espécies do grupo *obscura* (Bachmann e Sperlich, 1993). No grupo *melanogaster*, os satélites dodeca (10 pb) e 1.688 (359 pb) foram identificados e caracterizados como sequências abundantes nos genomas de várias espécies (Hsieh e Brutlag, 1979; Abad e col. 1992; Lohe e col. 1993). Já no grupo *repleta*, Kuhn e Sene (2005) identificaram a família pBuM com monômeros de 190 pb ou 370 pb em espécies do cluster *buzzatii.*

Embora várias famílias de DNAs satélites já tenham sido identificadas e caracterizadas em espécies de *Drosophila*, alguns grupos ainda permanecem pouco estudados, como por exemplo o grupo *montium*. Neste caso, a escassez de genomas sequenciados de espécies do grupo era, até pouco tempo, um fator limitante para trabalhos relacionados à genômica de elementos repetitivos. Porém, o recente sequenciamento dos genomas de várias espécies permite, agora, um acesso mais fácil, rápido e detalhado dos DNAs satélites deste grupo (Bronski e col. 2020; Conner e col. 2021). Aliado a isso, novas plataformas, softwares e pipelines têm sido desenvolvidos especialmente para a realização de análises *in silico* com o objetivo da identificação *de novo* e caracterização de sequências de DNAs satélites em espécies de eucariotos com genomas sequenciados via tecnologias NGS.

1.4. RepeatExplorer e TAREAN pipelines: identificação *de novo* de DNA satélites

Embora as tecnologias de sequenciamento de nova geração (NGS) venham revolucionando e aperfeiçoando os estudos genômicos, o processamento e montagem de *reads* curtas (<1kb) pós-sequenciamento ainda apresenta algumas limitações. Uma delas é a dificuldade na montagem de regiões dos genomas que são ricas em sequências repetitivas, como por exemplo os centrômeros (Grady e col. 1992; Sun e col. 2003; Plohl e col. 2014). Este problema, muitas vezes, implica na subamostragem ou até mesmo a ausência de sequências satélites e outros elementos repetitivos no genoma final montado, o que dificulta a identificação e caracterização correta de satDNAs. Uma forma de resolver este entrave é por meio da identificação destes DNAs repetitivos a partir de *reads* brutas não-montadas, geradas pós-sequenciamento de DNA. Esta abordagem foi implementada nos pipelines computacionais RepeatExplorex e TAREAN (Novák e col. 2013, 2017).

O RepeatExplorer e o TAREAN (*Tandem Repeat Analyzer*) são pipelines desenvolvidos para a identificação de elementos repetitivos no genoma de eucariotos. Enquanto o RepeatExplorer detecta diferentes tipos de elementos repetitivos no genoma fornecido, estando eles dispersos ou em tandem, o TAREAN identifica apenas elementos repetitivos organizados em tandem. Nos últimos anos, ambas as ferramentas vêm sendo utilizadas para o estudo de sequências repetitivas em várias espécies de eucariotos (García e col. 2015; Ruiz-Ruano e col. 2016; Robledillo e col. 2018; Sena e col. 2020; Valeri e col. 2020; Dias e col. 2021). No entanto, poucas espécies de *Drosophila* tiveram seu satelitoma determinado com essas novas abordagens *in silico* (de Lima e col. 2017).

Os pipelines RepeatExplorer e TAREAN requerem como *input* um arquivo FASTQ único formado por *single* ou *paired-end reads* gerados pós sequenciamento do genoma por abordagem *shotgun*. Geralmente, *paired-end* reads fornecem resultados melhores do que análises realizadas com *single reads* (Novák e col. 2017). Para iniciar as análises, é necessário que: as *reads* possuam tamanho uniforme (entre 100 e 200pb), que o número total de *reads* analisadas seja menor que a cobertura de 1x do genoma (as diretrizes do *pipeline* recomendam uma cobertura do genoma de 0,01 - 0,50 x), que as *reads* sejam pré-filtradas por qualidade (*score* de qualidade >= 10 em 95% das bases e sem presença de “Ns”) e que apenas *reads* completas sejam incluídas na análise.

Os *pipelines* acima descritos foram desenvolvidos com base na montagem de grafos de Bruijn (Compeau e col. 2011). Inicialmente, as *reads* são agrupadas em *clusters* de acordo com a similaridade das sequências e, em seguida, são reorientadas

para a formação de *k-mers* e, posteriormente, formação de sequências consensos. Para análises realizadas com o TAREAN, os grafos gerados também são pontuados em relação à presença de arranjos de cópias em tandem (Figura 5A). Além disso, clusters identificados pelo TAREAN são classificados como satélites de alta ou baixa confiança. Tais estimativas são definidas de acordo com os índices “Connected component index (C)” e “Pair completeness index (P)”. O índice “C” diz respeito a sequências genômicas repetidas em tandem, já o índice “P” mede a razão entre o número de *paired-end* reads completas e “quebradas”, o que está diretamente relacionado ao tamanho dessas cadeias contínuas de repetições em tandem, característica importante para a correta classificação de DNAs satélites (Novák e col. 2017). Desta maneira, quanto maiores os valores de C e P, mais circular será o grafo resultante e maior a probabilidade do cluster gerado corresponder a um satDNA verdadeiro (Figura 5C).

Embora a confiabilidade da classificação de “DNA satélite” seja um importante critério a ser levado em consideração para a identificação de satDNAs feitos pelo TAREAN, esta classificação deve ser feita com cautela. Por exemplo, alguns clusters classificados previamente como satDNAs de baixa confiança podem representar, na verdade, satDNAs verdadeiros (Figura 5B) (Novák e col. 2017). Isso acontece porque alguns clusters podem apresentar características intermediárias de sequências satélites, como por exemplo, valor C alto e valor P baixo, ou vice-versa. Desta maneira, a categorização e valores de limiar utilizados pelas *pipelines* devem ser interpretadas com cautela, uma vez que, levando em consideração os diferentes tipos de sequência que estão presentes nos genomas, faz se necessário uma investigação mais detalhada de cada cluster identificado *in silico*. Neste contexto, uma melhor classificação e caracterização de DNAs satélites deve também levar em consideração outros critérios, como localização cromossômica das sequências, o que pode ser obtido por meio de experimentos de citogenética molecular.

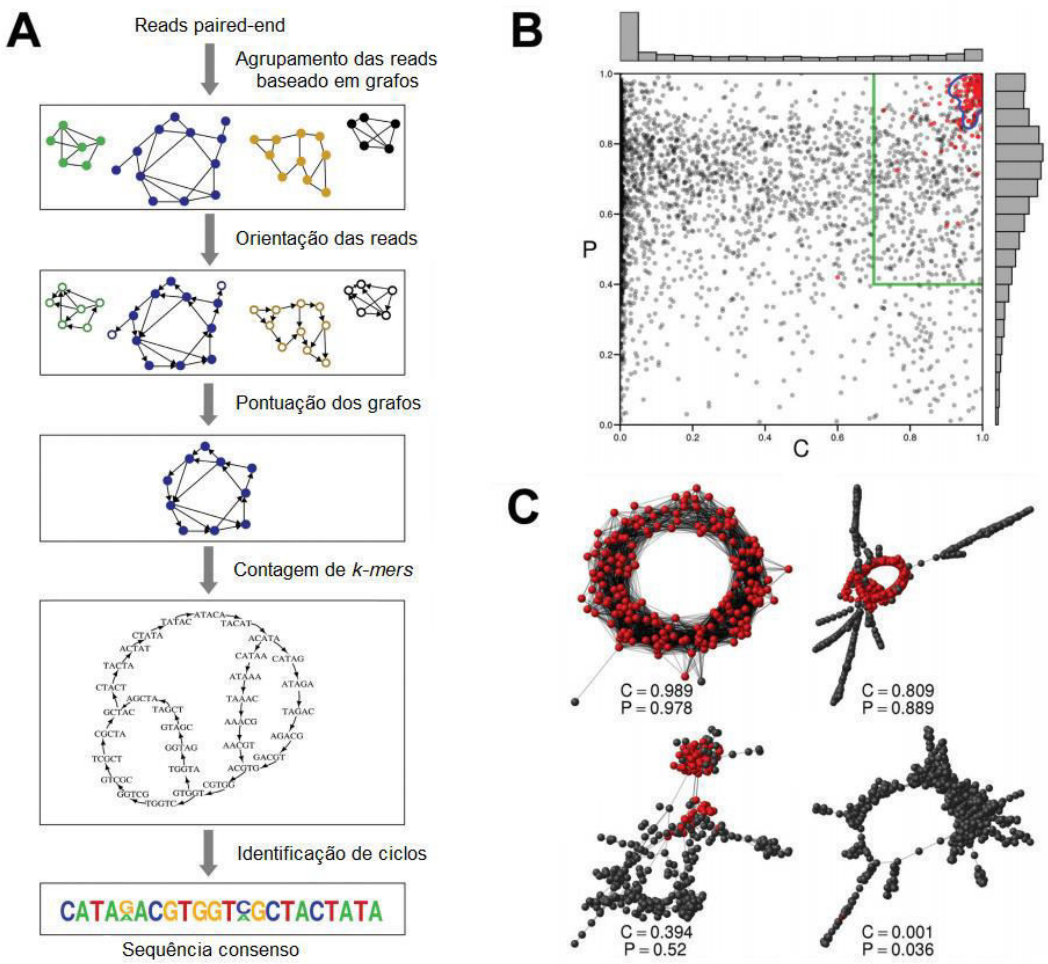


Figura 5. Pipeline TAREAN para identificação *de novo* de satDNAs. (Adaptado de Novák e col. 2017). **A**. *Workflow* de análises feitas no *pipeline* TAREAN. **B**. Exemplo de *scatter plot* de valores C e P para clusters identificados pelo TAREAN. Pontos vermelhos são clusters que foram anotados manualmente como satDNAs. A Linha azul indica o *threshold* definido para clusters identificados como satDNAs de alta confiança (*high confidence*), enquanto que a linha verde indica o *threshold* definido para clusters identificados como satDNAs de baixa confiança (*low confidence*). Note a presença de clusters que foram classificados como satélites de baixa confiança pelo TAREAN mas que, após anotação manual, foram considerados satDNAs verdadeiros. **C**. Exemplos de clusters visualizados como grafos, onde os nós representam as reads e os vértices as *reads* conectadas por sequências similares. Os nós pertencentes às sequências mais fortemente conectadas dentro dos grafos são mostrados em vermelho. Valores C e P correspondentes são mostrados abaixo de cada grafo.

2. Objetivo

O objetivo principal desse projeto foi realizar a identificação *de novo* de satDNAs com os pipelines RepeatExplorer e TAREAN em genomas de duas espécies de *Drosophila* do grupo *virilis* (*D. virilis* e *D. americana*) e em 23 espécies recém-sequenciadas do grupo *montium*.

2.1. Objetivos específicos

2.1.1. Analisar a capacidade dos pipelines RepeatExplorer e TAREAN na identificação dos DNAs satélites mais abundantes de *D. virilis* e *D. americana*;

2.1.2. Mapear os prováveis DNAs satélites mais abundantes identificados pelo RepeatExplorer e TAREAN em cromossomos metafásicos e politênicos de *D. virilis* e *D. americana*;

2.1.3. Identificar com o pipeline TAREAN prováveis satDNAs nos 23 genomas recém-sequenciados de espécies do grupo *montium*;

2.1.4. Avaliar se as famílias de satDNAs identificadas nas espécies do grupo *montium* são bons marcadores taxonômicos e filogenéticos para o grupo.

Os principais resultados do projeto, bem como discussão dos mesmos, serão apresentados na forma de dois capítulos, elaborados na forma de artigos. O primeiro capítulo foi publicado em dezembro de 2019 na revista PLOS One e apresenta os resultados obtidos para as análises realizadas em *D. virilis* e *D. americana*, espécies do grupo *virilis*. O segundo capítulo, ainda não publicado, apresenta os resultados obtidos dos DNAs satélites identificados nas espécies do grupo *montium*.

3. Capítulo 1: Artigo "*De novo* identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines"

PLOS ONE

RESEARCH ARTICLE

De novo identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines

Bráulio S. M. L. Silva, Pedro Heringer, Guilherme B. Dias ¤, Marta Svartman, Gustavo C. S. Kuhn*

Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil

¤ Current address: Department of Genetics and Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America
* gcskuhn@ufmg.br

OPEN ACCESS

Citation: Silva BSML, Heringer P, Dias GB, Svartman M, Kuhn GCS (2019) *De novo* identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines. PLoS ONE 14(12): e0223466. https://doi.org/10.1371/journal.pone.0223466

Editor: Ruslan Kalendar, University of Helsinki, FINLAND

Received: September 19, 2019

Accepted: November 26, 2019

Published: December 19, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by "Conselho Nacional de Desenvolvimento Cient´ifico e Tecnológico" (CNPq - http://www.cnpq.br/) to G.K. (Grant: 404620/2016-7 and Fellowship: 308386/2018-3) and to M.S. (310433/2018-5), and a fellowship from "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"

Abstract

Satellite DNAs are among the most abundant repetitive DNAs found in eukaryote genomes, where they participate in a variety of biological roles, from being components of important chromosome structures to gene regulation. Experimental methodologies used before the genomic era were insufficient, too laborious and time-consuming to recover the collection of all satDNAs from a genome. Today, the availability of whole sequenced genomes combined with the development of specific bioinformatic tools are expected to foster the identification of virtually all the "satellitome" of a particular species. While whole genome assemblies are important to obtain a global view of genome organization, most of them are incomplete and lack repetitive regions. We applied short-read sequencing and similarity clustering in order to perform a *de novo* identification of the most abundant satellite families in two *Drosophila* species from the *virilis* group: *Drosophila virilis* and *D. americana*, using the Tandem Repeat Analyzer (TAREAN) and RepeatExplorer pipelines. These species were chosen because they have been used as models to understand satDNA biology since the early 70's. We combined the computational approach with data from the literature and chromosome mapping to obtain an overview of the major tandem repeat sequences of these species. The fact that all of the abundant tandem repeats (TRs) we detected were previously identified in the literature allowed us to evaluate the efficiency of TAREAN in correctly identifying true satDNAs. Our results indicate that raw sequencing reads can be efficiently used to detect satDNAs, but that abundant tandem repeats present in dispersed arrays or associated with transposable elements are frequent false positives. We demonstrate that TAREAN with its parent method RepeatExplorer may be used as resources to detect tandem repeats associated with transposable elements and also to reveal families of dispersed tandem repeats.

(CAPES - https://www.capes.gov.br/) to B.S., P.H. and G.D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The genome of eukaryotes encloses a variety of repetitive DNA sequences which comprises most of the nuclear DNA of several organisms, including animals, plants and insects [1,2]. Among them are the satellite DNAs (satDNAs), usually defined as abundant, tandemly repeated noncoding DNA sequences, forming large arrays (hundreds of kilobases up to megabases), typically located in the heterochromatic regions of the chromosomes [3,4], although short arrays may additionally be present in the euchromatin [5,6].

The collection of satDNAs in the genome, also known as the "satellitome", usually represents a significant fraction (>30%) of several animal and plant genomes. Other classes of noncoding tandem repeats include the microsatellites, with repeat units less than 10 bp long, array sizes around 100 bp and scattered throughout the genome; and the minisatellites, with repeats between 10 to 100 bp long, forming up to kb-size arrays, located at several euchromatic regions, with a high density at terminal chromosome regions [3,4]. Therefore, the best criteria to distinguish satellites from micro and minisatellites are long array sizes and preferential accumulation at heterochromatin for the former.

SatDNAs do not encode proteins, but they may play important functional roles in the chromosomes, most notably related to chromatin modulation and the establishment of centromeres [7–9]. They are among the fastest evolving components of the genome (although some conserved satellites have also been reported) [10–12], and such behavior combined to their abundance and structural role have major implications for the evolution and diversification of genomes and species [8,13].

Since the discovery of satDNAs in the early 60's, species from the genus *Drosophila* have been used as a model to address several aspects of satDNA biology, such as their origin, organization, variation, evolution and function (e.g. [7,14–18]).

Currently, several *Drosophila* genomes have been sequenced by next-generation technologies and new bioinformatic tools have been designed for the identification of repetitive DNAs from this vast source of genomic resources [19]. Among them, the RepeatExplorer software [20] has been successfully used for *de novo* identification of repetitive DNAs directly from unassembled short sequence reads, and the recently implemented Tandem Repeat Analyzer (TAREAN) pipeline [21] was introduced to specifically identify putative satDNAs. Such a combination between sequenced genomes and bioinformatic tools is now expected to foster the identification of the full "satellitome" of any given species (e.g. [22–26]). Despite the availability of all such resources, only a few *Drosophila* species had their satDNA landscape determined with these new approaches [23].

In the genus *Drosophila* genome sizes vary between ~130 Mb to ~400 Mb, but most analyzed species have genome with around 180–200 Mb, such as *D. melanogaster* [27,28]. The satDNA content also varies across species, from ~2% in *D. buzzatii* [23] to ~60% in *D. nasutoides* [29]. Some studies suggest a positive correlation between genome size and the amount of satDNAs in *Drosophila* [28,30,31].

The genome size of *D. virilis* (*virilis* group), with ~400 Mb, is among the largest reported for *Drosophila*. Accordingly, the estimated satDNA in this species is also high (>40%) [28,32]. Previous studies using CsCl density gradients revealed that three evolutionary related satDNAs with 7 bp long repeat units and only one mutation difference, named satellite1 (5' ACAAACT 3'), satellite2 (5' ATAAACT 3') and satellite3 (5' ACAAATT 3') together represent ~40% of its genome [32,33]. These satellites mapped predominantly to the heterochromatic regions of all chromosomes except the Y. Another satDNA identified in this species, but using genomic DNA digestion with restriction endonucleases, was named pvB370, and consists of 370 bp long repeat units [34] predominantly located at sub-telomeric regions and, to a lesser extent,

along some discrete euchromatic loci [35]. Other abundant TRs have been identified in the *D. virilis* genome, such as the 220TR and 154TR families, which belong to the internal structure of transposable elements [16,36], the 225 bp family, present in the intergenic spacer of ribosomal genes, and the less characterized 172 bp family [37]. A recent study reported additional tandem repeats less than 20 bp long but at low abundance [18].

The high throughput and low cost of current whole-genome sequencing technologies have made it possible to obtain genome assemblies for a wide range of organisms. However, *de novo* whole-genome shotgun strategies are still largely unable to fully recover highly repetitive regions such as centromeres and pericentromeric regions and, as a result, satDNAs are usually misrepresented or absent from such assemblies [19]. One way of circumventing the assembly bottleneck is to directly identify repeats from raw sequencing reads. One of such approaches is implemented in the RepeatExplorer pipeline, already used in a wide range of plant and animal species [22,38,39]. RepeatExplorer performs similarity-based clustering of raw short sequencing reads and partial consensus assembly, allowing for repeat identification even from small samples of genome coverage. A recent development of RepeatExplorer includes the TAREAN pipeline for the specific detection of tandem repeats by searching for circular structures in directed read clusters [21].

In the present study, we aimed to test the ability of TAREAN to correctly identify and estimate the abundance of satDNAs in *D. virilis*. To refine and expand our knowledge of the identified putative satDNAs, in some cases we mapped them in mitotic and polytene chromosomes using fluorescent in situ hybridization (FISH) technique.

There are several examples showing that satDNA abundance may vary widely even across closely related species [10,40]. For example, one species may present few repeats in the genome (therefore not being identified as a satellite), while a closely related species presents thousands, reaching a satDNA status. For this reason, we also added to our study *D. americana*, a species belonging from the *virilis* group, but separated from *D. virilis* by ~4.1 Myr [41].

Material and methods

RepeatExplorer and TAREAN analyses

The *in silico* identification of putative satDNAs was performed using the RepeatExplorer and TAREAN pipelines [20,21] implemented in the Galaxy platform [42]. These algorithms were developed to identify and characterize repetitive DNA elements from unassembled short read sequences. We used the publicly available *Drosophila virilis* strain 160 (SRX669289), *Drosophila americana* strain H5 (ERX1035147) and *Drosophila americana* strain W11 (ERX1035149) [43] Illumina paired-end sequences. The sequences were obtained through the "European Nucleotide Archive" (EBI) database and their quality scores measured with the "FASTQC" tool. We used "FASTQ Groomer" (Sanger & Illumina 1.8 +) to convert all the sequences to a single fastqsanger format. We removed adapters and excluded any reads with more than 5% of its sequence in low quality bases (Phred cutoff < 10) using the "Preprocessing of fastq paired-reads" tool included in the RepeatExplorer Galaxy instance. The interlaced filtered paired-end reads were used as input data for the RepeatExplorer clustering and Tandem Repeat Analyzer tools with the following settings: "sample size = 2,000,000—select taxon and protein domain database version (REXdb): Metazoa version 3.0—select queue: extra-long and slow". For the TAREAN analyses we also used the "perform cluster merging" tool for reducing the redundancy of the results.

The results were provided in a HTML archive report and all the data were downloaded in a single archive for further investigation. We analyzed clusters representing $>0.5\%$ of the genome of *Drosophila virilis* strain 160.

Clusters with tandem repeats identified by TAREAN are denoted as putative high or low confidence satellites. These estimates are denoted according to the "Connected component index (C)" and "Pair completeness index (P)". The C index indicates clusters formed by tandemly repeated genomic sequences, while the P index measures the ratio between complete read pairs in the cluster and the number of broken pairs, that is directly related to the length of continuous tandem arrays [21].

Fluorescent probe construction

We extracted total genomic DNA from a pool of 20 adult *Drosophila virilis* (strain 15010–1051.51 from Santiago, Chile) and *D. americana* (strain H5 from Mississipi, United States of America) with the Wizard[1] Genomic DNA Purification Kit (Promega Corporation). For primer's design, we used the consensus sequences from each satDNA identified by RepeatExplorer/TAREAN and multiple sequence alignments by selecting the most conserved nucleotide regions. Satellite DNAs were PCR amplified with the following primers forward (F) and reverse (R):

Sat1_F (ACAAACTACAAACTACAAACTACAAACTACAAACT), Sat1_R (AGTTTGTAG TTTGTAGTTTGTAGTTTGTAGTTTGT), 172TR_F (ATTTATGGGCTGGGAAGCTTTGA CGTATG), 172TR_R (CGGTCAAATCTCATCCGATTTTCATGAGG), 225TR_F (GCGACACCACTCCCTATATAGG), 225TR_R (CGCGCAAGGCATGTCATATG), pvB370_F (TAGTAGGGATCCGTACAAATTCAA), pvB370_R (GTACGGATCCCTACTAATAATTGGCAT).

All primers were used to amplify the target sequences from genomic DNA, with the exception of Sat1 in which the amplification process was conducted by forward and reverse primers self-annealing without genomic DNA. The PCR products were excised from agarose gels and ligated into pGEM-T vector plasmids (Promega) with T4 DNA ligase (Promega). For cloning, the plasmids were multiplied into *E.coli* cells and then eluted with the PureLink™ Quick Plasmid Miniprep Kit (Invitrogen). To ensure the presence of the inserts, the final samples were Sanger sequenced in an ABI3130 and later analyzed in the Chromas software (Technelysium). Clones with satDNA inserts were later prepared as probes for FISH.

Fluorescent in situ hybridization (FISH)

The metaphase and polytene chromosomes were obtained from neuroblasts and salivary glands of third instar larvae of *D. virilis* (strain 15010–1051.51) and *D. americana* (strain H5), according to [44,45]. Probe labeling and FISH experiment conditions were conducted according to [16]. The satDNA probes were immunodetected with antidigoxigenin-Rhodamine and avidin-FITC (Roche Applied Science).

We used DAPI "4,6-diamidino-2-phenylindole" (Roche) in "SlowFade" antifade reagent (Invitrogen) for DNA counterstaining. The analyses were conducted under an Axio Imager A2 epifluorescence microscope equipped with the AxiocamMRm camera (Zeiss). Images were captured with Axiovision (Zeiss) and edited in Adobe Photoshop.

Results

Identification of putative satDNAs in *D. virilis* and *D. americana*

The most abundant putative satDNAs (covering >0.5% of the genome) identified by the RepeatExplorer and TAREAN pipelines are shown in Table 1 (see S1 Fig for histogram summary analyses and S4–S15 Figs for detailed data from each cluster retrieved). All of the six

Table 1. Putative satellite DNAs in *D. virilis* strain 160 and *D. americana* strain H5 identified by TAREAN and Repeat Explorer.

	Drosophila virilis 160						*Drosophila americana* H5					
Tandem repeat family[a]	**Sat1**	**154TR**	**pvB370**	**172TR**	**225TR**	**36TR**	**Sat1**	**172TR**	**154TR**	**pvB370**	**225TR**	**36TR**
Satellite confidence	High	Low	Low	High	Low	Low□	High	Low	Low	High	High□	n/a□
Satellite probability	0.92	0.03	0.53	0.73	0.69	0.00□	0.91	0.69	0.04	0.75	0.76□	0.00□
C index[b]	0.98	0.94	0.96	0.97	0.99	0.94□	0.96	0.97	0.94	0.97	0.97□	0.72□
P index[b]	0.92	0.71	0.81	0.86	0.87	0.52□	0.97	0.85	0.72	0.87	0.86□	0.24□
Consensus size	7bp	154bp	370bp	171bp	225bp	36bp□	7bp	171bp	154bp	199bp	225bp□	n/a□
Genome proportion (%)	12.0	1.6	1.6	1.1	0.8	0.7□	9.0	2.7	2.2	1.7	0.9□	0.4□

□. Results obtained from RepeatExplorer instead of TAREAN.

[a]. Ordered by abundance from higher to lower.

[b]. C and P indexes are explained in Materials and Methods.

https://doi.org/10.1371/journal.pone.0223466.t001

identified tandem repeat families (Sat1, 154TR, pvB370, 172TR, 225TR, 36TR) are shared by both species and have been previously identified.

Although the total abundance of these six tandem repeats is similar (~17%) in the two species, there are differences in the estimated proportion occupied by each putative satDNA between the species.

In order to check if these differences are predominantly inter-specific, we used RepeatExplorer and TAREAN to compare the abundances of each tandem repeat between two *D. americana* strains (H5 and W11), which were sequenced using the same sequencing platform and methods. Our analysis showed that differences in repeat proportion among *D. americana* strains are somewhat comparable with the ones observed between *D. virilis* and *D. americana* (Fig 1). These results indicate that comparisons of tandem repeat abundance between taxa using RepeatExplorer and TAREAN should be taken with caution as significant differences can also be observed among lineages within the same species. Interestingly, repeat abundance variations between lineages in these species were also detected by [46].

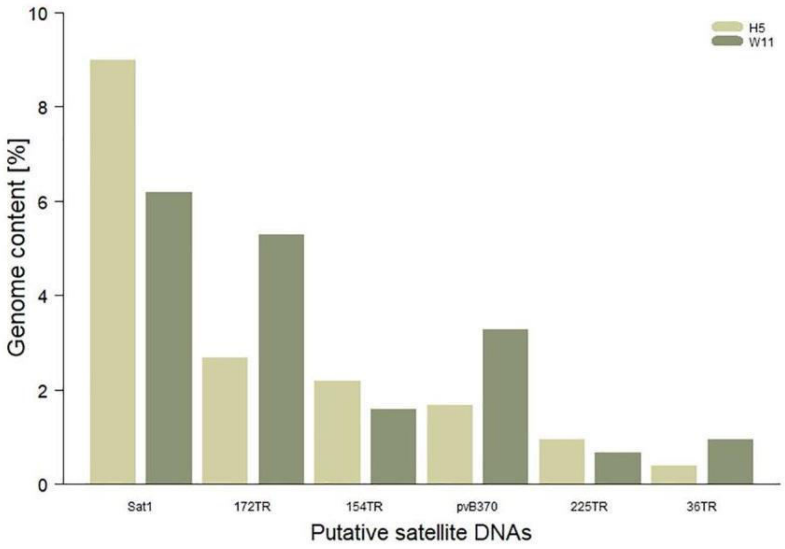


Fig 1. Genome content for six putative satellite DNAs in two *Drosophila americana* strains (H5 and W11) according to RepeatExplorer and TAREAN analyses.

https://doi.org/10.1371/journal.pone.0223466.g001

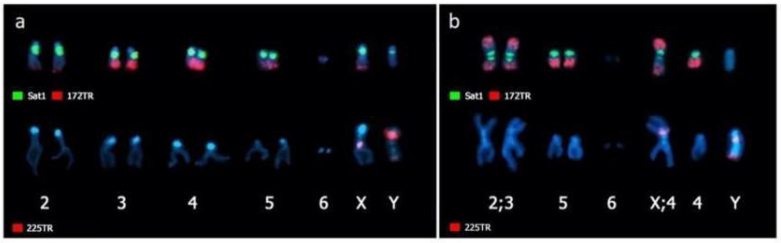


Fig 2. Mapping of Sat1, 172TR and 225TR by FISH on metaphase chromosomes. (A) *Drosophila virilis* and (B) *Drosophila americana*. Upper panel: Sat1 (green) and 172TR (red). Lower panel: 225TR (red). The mitotic chromosomes of *D. virilis* were identified by their sizes combined with the hybridization signals on polytene chromosomes (see Fig 3).

https://doi.org/10.1371/journal.pone.0223466.g002

To further characterize the tandem repeat families identified *in silico*, we constructed DNA probes using the consensus sequences generated by TAREAN from three families and used them to verify their localization in metaphase and polytene chromosomes. In the following sections we describe our *in silico* and FISH analyses for each identified family, comparing the results with previous studies and discussing if TAREAN correctly identified and distinguished satDNAs from other classes of tandem repeats. The tandem repeat families are described below in order of their abundance (higher to lower) as revealed for *D. virilis* strain 160.

Sat1

The most abundant tandem repeat identified by TAREAN in *D. virilis* and *D. americana* is composed by a 7 bp long repeat corresponding to the previous described satellite I [33]. In *D. virilis*, our FISH experiments in metaphase chromosomes showed this satDNA occupying the pericentromeric region of all autosomes except the small dot chromosomes, and in the X and Y chromosomes (Fig 2A). However, the hybridization in polytene chromosomes revealed that Sat1 also localizes in the pericentromeric region of the dot chromosome (Fig 3A). [32] showed a similar hybridization pattern, although their results did not consistently demonstrate Sat1 signals in the dot and Y chromosomes.

In *D. americana*, Sat1 signals were detected in the pericentromeric region of all autosomes in metaphase chromosomes, except the dot (Fig 2B), while in polytene chromosomes, Sat1 signals were also observed in the dot chromosomes (Fig 3B). However, differently to what was observed in *D. virilis*, our Sat1 hybridizations in the *D. americana* polytene dot chromosomes did not give enough information about the precise location of this satDNA, although it also appears to occupy a portion of the pericentromeric region. As another difference from *D. virilis*, Sat1 sequences appear to be absent from the Y chromosome in *D. americana* (Fig 2B). Our FISH results corroborate the smaller genomic fraction occupied by this satDNA in *D. americana* (~9% against ~12% in *D. virilis*), revealed by the *in silico* analysis (Figs 2, 3A and 3B). These new findings in *D. americana* and *D. virilis* also agree with recent results from [46].

154TR

The genomic distribution of 154TR has been recently studied in detail in *D. virilis* and *D. americana* using FISH in metaphase and polytene chromosomes [36]. This sequence was independently identified *in silico* by [37] and [48]. The 154TR was characterized as a tandem repeat derived from a Helitron transposable element [37], which was studied in detail and classified as a family named DINE-TR1 [36]. DINE-TR1 elements containing 154TR homologous sequences were found in several *Acalyptratae* species, mostly within the *Drosophila* genus,

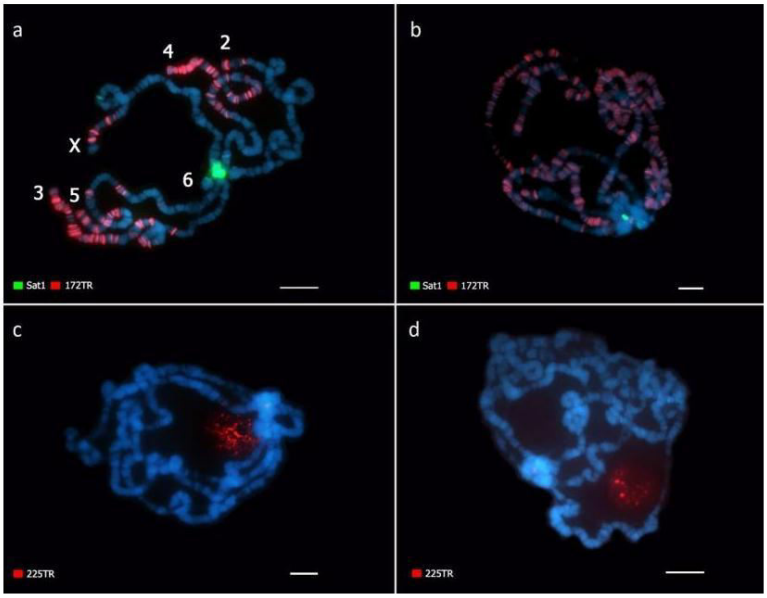


Fig 3. Mapping of Sat1, 172TR and 225TR by FISH on polytene chromosomes. (A, C) *Drosophila virilis* and (B, D) *Drosophila americana*. (A, B) Sat1 (green) and 172TR (red). (C, D) 225TR (red). Scale bars represent 10μm. The polytene chromosome arms were identified according to [47].

https://doi.org/10.1371/journal.pone.0223466.g003

although long arrays (> 10 copies) of 154TR were only detected in three species (*D. virilis*, *D. americana* and *D. biarmipes*) [36].

FISH in metaphase and polytene chromosomes revealed that 154TR is located in the distal pericentromeric region (β-heterochromatin) and many euchromatic loci of all autosomes and the X chromosome of *D. virilis* and *D. americana*. In addition, this tandem repeat covers a large portion of the Y chromosome in both species. In *D. virilis*, 154TR signals are very abundant in the centromeric heterochromatin of chromosome 5 and are also found in a discrete region within the pericentromeric region of the X chromosome [36].

Our results from the TAREAN analysis classified 154TR as a putative satellite with low confidence in both species (Table 1). We suggest that this result is probably a consequence of 154TR being both tandemly repeated, like a satDNA, and dispersed, like a transposable element. In this case, even though the connected component index (C) of 154TR is high, its relatively low pair completeness index (P) contributes to its classification as a putative satellite with low confidence by TAREAN (Table 1). We suggest that 154TR is not a satDNA and thus, should be classified as a highly abundant dispersed tandem repeat.

pvB370

The pvB370 satellite was first described by [34], who also identified this family as deriving from the direct terminal repeats of pDv transposable elements [49]. In a following study, [35] showed that in *D. virilis* and *D. americana*, pvB370 is located at several euchromatic loci and at the telomeric region of all chromosomes.

Because pvB370 was previously mapped in the chromosomes of *D. virilis* and *D. americana* using FISH, we did not conduct a throughout analysis on both species. However, because pvB370 seems to display a euchromatic distribution [35] similar to the one we observed for 172TR (Fig 3A and 3B) we hybridized both pvB370 and 172TR probes concomitantly in *D. americana* polytene chromosomes. Our results showed little or no overlap between pvB370

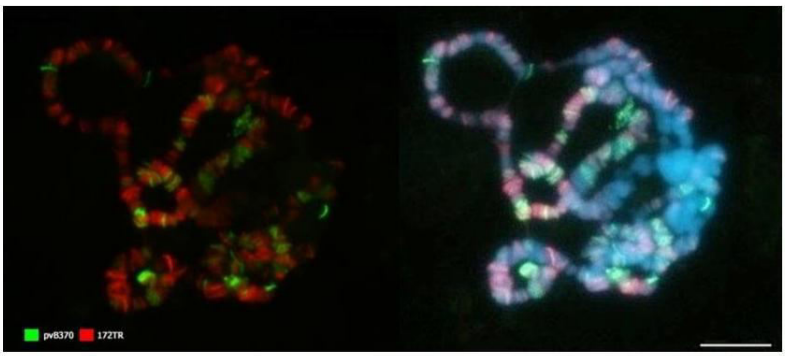


Fig 4. Chromosome location of 172TR and pvB370 by FISH on polytene chromosomes of *Drosophila americana*. There is little or no overlap between these tandem repeats. Red (172TR) and green (pvB370). Scale bar represents 10μm.

https://doi.org/10.1371/journal.pone.0223466.g004

and 172TR, although many arrays from the two families are very close (at least a few kbp) to each other (Fig 4).

172TR

The 172TR family corresponds to the 172 bp tandem repeats previously identified *in silico* by [37]. Our FISH results in the metaphase and polytene chromosomes of *D. virilis* revealed that 172TR is distributed throughout the arms of autosomes 3, 4 and 5, in several loci at the X chromosome and in at least two loci in chromosome 2, including the subtelomeric region (Figs 2A and 3A). Most of the arrays are located at distal chromosome regions. No hybridization signals were detected in the dot and Y chromosomes.

The FISH results in *D. americana* showed 172TR signals at multiple loci along all autosomes, except the dot, and more equally distributed in both distal and proximal regions of chromosome arms (Figs 2B and 3B). Similarly to *D. virilis*, no hybridization signal was detected in the Y chromosome (Fig 2B). The FISH data (Figs 2, 3A and 3B) clearly showed a higher number of 172TR loci in *D. americana* compared to *D. virilis*, a result that is consistent with the higher overall abundance of 172TR repeats in *D. americana* predicted by the *in silico* analysis (Table 1).

225TR

The putative satDNA detected in our *in silico* analyses as 225TR was previously identified as a component of intergenic spacers (IGS) of ribosomal genes from *D. virilis* located at the chromocenter and nucleolus regions of polytene chromosomes [37]. Our FISH experiments in polytene chromosomes confirmed these results in *D. virilis* (Fig 3C), additionally showing that in *D. americana* this family displays the same pattern of localization (Fig 3D).

In addition, we also performed FISH with a 225TR probe in metaphase chromosomes of both species for the first time, that revealed its location in the pericentromeric region of the X chromosome and in the pericentromeric and telomeric regions of chromosome Y (Fig 2A and 2B). This result is in accordance with previous studies showing the location of these IGS sequences in the sex chromosomes of *Drosophila* [50].

Although the TAREAN pipeline failed to detect the 225TR in *D. americana*, RepeatExplorer revealed the presence of this family. This indicates a possible limitation of TAREAN in detecting less abundant tandem repeats in comparison with RepeatExplorer. Moreover, TAREAN only retrieves clusters with highly circular structures, and therefore excludes 225TR repeats

that are associated with linear structures (S12 and S13 Figs). These observations indicate that, although 225TR is an abundant tandem repeat, it does not have all the typical features of a satDNA.

36TR

A previous work made by [49] identified the presence of 36 bp tandem repeats inside the pDv transposable element and a subsequent work by [34] showed that array size variation exists among different pDv copies in *D. virilis*. This TR was not retrieved by the TAREAN pipeline but we found it in high abundance (~0.73% in *D. virilis* and ~0.48% in *D. americana*; Table 1) among the results from RepeatExplorer, that further classified this TR as a low confidence satDNA. Interestingly, the RepeatExplorer pipeline revealed that the cluster corresponding to this 36 bp tandem repeat has a high number of shared reads with the pvB370 cluster (S2 and S3 Figs). In this case, the link between 36TR and pvB370 clusters is explained by their co-occurrence as complete (36 bp) and partial (pvB370) sequences within the pDv transposable element [34]. This result shows that the RepeatExplorer pipeline is able to detect putative relationships between distinct repetitive sequences.

Discussion

Here we performed *de novo* identification of the most abundant tandem repeat families in *D. virilis* and *D. americana*. These species were chosen because they have larger genomes compared to other *Drosophila* species and because they have been used as models to understand satDNA biology since the early 70's. In order to do that, we combined the RepeatExplorer and TAREAN results with data from the literature and, in some cases, with new chromosome mapping data obtained by us using FISH in metaphase and polytene chromosomes.

Because all of the repeats identified herein had been previously detected by other methods, we were able to test if the TAREAN pipeline could correctly classify them as satDNAs or not.

TAREAN identified the heptanucleotide Sat1 as a satDNA with high confidence, which agrees with all attributes known for this family and the satDNA definition (i.e. high copy-number, long-arrays, predominant heterochromatic location) [32,33]. Sat1 was identified as the most abundant tandem repeat in both *D. virilis* and *D. americana*, which is also in accordance with previous work [32,33]. However, the other two less abundant heptanucleotide satellites, Sat2 and Sat3, were not detected by TAREAN. As these three satellites differ from each other by a single nucleotide substitution, they were likely all included in the Sat1 cluster by TAREAN. This clustering of variants appears to be a relevant disadvantage that might influence the identification of not only the heptanucleotide satDNA family but other short repeat families with similar features (e.g. short monomer size and high sequence similarity). Therefore, to analyze these type of sequences in detail, it might be advisable to also use tools that are more appropriate for this aim, for example, the software k-Seek [51]. It is also worth mentioning that the heptanucleotide satDNA genomic fractions revealed by TAREAN (~12% for *D. virilis* strain 160 and ~9% for *D. americana* strain H5) are significantly below the previously estimated of >40% genomic fraction, based on density gradient ultracentrifugation methods [32,52]. Although TAREAN may not be ideally suitable to quantify satellites with short repeat units [21], it is worth mentioning that [46] have recently demonstrated that Illumina sequence reads containing the heptanucleotide satellites from *D. virilis* tend to be highly enriched for low quality scores. Furthermore, the use of raw reads from different sequencing platforms did not allowed the recovery of simple satellites at the predicted ~40% genomic fraction indicated by previous works [46]. The difference between these estimates (12% to 40%) may reflect an intrinsic bias in current sequencing methods. A second possibility, which does not reject the

first is the existence of real differences in satDNA content between different strains of the same species.

TAREAN classified the 154TR, pvB370 and 36TR families as putative satellites in *D. virilis* and *D. americana*. With the exception of pvB370 in *D. americana*, which was classified with high confidence, all remaining repeats had low confidence calls from TAREAN (Table 1). These tandem repeats are known to be abundant and associated with transposable elements (as integral parts or evolutionarily related), suggesting that RepeatExplorer and TAREAN could be used as resources to detect tandem repeats associated with transposable elements. In the case of 154TR, pvB370 and 36TR, the relationship could be checked directly in the RepeatExplorer pipeline by identifying clusters of tandem repeats sharing a high number of reads with clusters associated to transposable elements (see S2 and S3 Figs), or indirectly in the TAREAN pipeline, by investigating the tandem repeats classified as putative satellites with low confidence (or lower values of satellite probability). The rationale behind this last procedure is that identified families with a 'low satellite score' may represent repetitive DNAs with intermediate features, being both highly dispersed and tandemly repeated. One situation in which this scenario is expected is the case where tandem repeats belonging to the terminal or internal portions of transposable elements underwent array expansion [36,53]. Nonetheless, some highly dispersed tandem repeats are not necessarily associated with transposable elements, which is the case of 172TR shown here and the 1.688 satDNA from *D. melanogaster* [5].

It is interesting to note that, in *D. virilis* and *D. americana*, the families 172TR, pvB370 and 154TR were either classified as putative satellites with low confidence, or with high confidence but associated with a relatively low satellite probability (Table 1). Because all these three families were found distributed along the euchromatic regions of chromosomes, we suggest that a low 'satellite score' in the TAREAN pipeline is a good predictor of dispersed tandem repeats. As mentioned above, although there is no indication of a relationship between the 172TR family with any known transposable element, its lower satellite score from the *in silico* analysis correctly predicts the dispersed array distribution observed in polytene chromosomes (Fig 3A and 3B).

In conclusion, six abundant putative satDNAs were identified in *D. virilis* and *D. americana* by TAREAN and RepeatExplorer: Sat1, 154TR, pvB370, 172TR, 225TR and 36TR. All of them have been previously characterized to a higher or lesser extent in previous works, but using different methodologies. The main advantage of TAREAN and RepeatExplorer in comparison with previous methods aiming to identify satDNAs in *D. virilis* refers to their relative lack of bias compared to the *in silico* digestion applied by [37], that identifies only tandem repeats presenting restriction sites, and the k-Seek method [51] applied by [18] that specifically identifies short tandem repeats with less than 20 bp.

While Sat1 (identified by TAREAN as a satDNA with high confidence) is in fact a family that matches all features typically attributed for satDNAs, the classification of the other families as satDNAs (identified as a satDNA with low confidence on at least one species) is more controversial. The 154TR, pvB370 and 36TR families are associated with the internal structure of TEs, thus being distributed along the chromosome arms with different degrees of dispersion. The 225TR belongs to the IGS of ribosomal genes. In contrast, the 172TR family is an abundant tandem repeat but with exclusive euchromatic location, where they apparently do not to reach satDNA-like long arrays. Based on the repeat unit length of 172TR (172 bp), this family cannot be considered as a micro or minisatellite. In this context, it would be interesting to further investigate these five families (154TR, pvB370, 172TR, 225TR and 36TR) using long-read sequencing technologies, since they are expected to provide more detailed information about their copy number and array sizes.

Supporting information

S1 Fig. TAREAN histogram summary analyses of (A) *Drosophila virilis* (strain 160) and (B) *Drosophila americana* (strain H5). The histogram analysis is the overall result of the clustering process, after filtering and pre-processing of raw reads. It shows (on the top), the total number of reads analyzed during the run. Each column represents a cluster (by abundance from left to right). The y-axis refers to the number of reads by cluster and the x-axis the percentage of each cluster in the analysis.
(PDF)

S2 Fig. pvB370 and 36TR supercluster analysis in *Drosophila virilis* strain 160.
(PDF)

S3 Fig. pvB370 and 36TR supercluster analysis in *Drosophila americana* strain H5.
(PDF)

S4 Fig. Sat1 cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S5 Fig. Sat1 cluster analysis in *Drosophila americana* strain H5.
(PDF)

S6 Fig. 154TR cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S7 Fig. 154TR cluster analysis in *Drosophila americana* strain H5.
(PDF)

S8 Fig. pvB370 cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S9 Fig. pvB370 cluster analysis in *Drosophila americana* strain H5.
(PDF)

S10 Fig. 172TR cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S11 Fig. 172TR cluster analysis in *Drosophila americana* strain H5.
(PDF)

S12 Fig. 225TR cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S13 Fig. 225TR cluster analysis in *Drosophila americana* strain H5.
(PDF)

S14 Fig. 36TR cluster analysis in *Drosophila virilis* strain 160.
(PDF)

S15 Fig. 36TR cluster analysis in Drosophila americana strain H5.
(PDF)

Acknowledgments

The authors wish to thank the two anonymous reviewers for their valuable comments and suggestions.

Author Contributions

Conceptualization: Bráulio S. M. L. Silva, Guilherme B. Dias, Gustavo C. S. Kuhn.

Data curation: Bráulio S. M. L. Silva, Pedro Heringer.

Formal analysis: Bráulio S. M. L. Silva, Pedro Heringer.

Funding acquisition: Marta Svartman, Gustavo C. S. Kuhn.

Investigation: Bráulio S. M. L. Silva, Pedro Heringer, Guilherme B. Dias, Gustavo C. S. Kuhn.

Methodology: Bráulio S. M. L. Silva, Guilherme B. Dias.

Project administration: Gustavo C. S. Kuhn.

Resources: Marta Svartman, Gustavo C. S. Kuhn.

Supervision: Gustavo C. S. Kuhn.

Validation: Gustavo C. S. Kuhn.

Writing – original draft: Bráulio S. M. L. Silva, Pedro Heringer.

Writing – review & editing: Bráulio S. M. L. Silva, Pedro Heringer, Guilherme B. Dias, Marta Svartman, Gustavo C. S. Kuhn.

References

1. de Koning APJ, Gu WJ, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two thirds of the human genome. PLoS genetics. 2011; 7(12). https://doi.org/10.1371/journal.pgen.1002384 PMID: 22144907
2. Biscotti MA, Olmo E, Heslop-Harrison JS. Repetitive DNA in eukaryotic genomes. Chromosome Research. 2015; 23(3):415–20. https://doi.org/10.1007/s10577-015-9499-z PMID: 26514350
3. Tautz D. Notes on the fefinition and nomenclature of tandemly repetitive DNA sequences. Exs. 1993; 67:21–8. https://doi.org/10.1007/978-3-0348-8583-6_2 PMID: 8400689
4. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994; 371(6494):215–20. https://doi.org/10.1038/371215a0 PMID: 8078581
5. Kuhn GCS, Kuttler H, Moreira O, Heslop-Harrison JS. The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. Mol Biol Evol. 2012; 29(1):7–11. https://doi.org/10.1093/molbev/msr173 PMID: 21712468
6. Pavlek M, Gelfand Y, Plohl M, Mestrovic N. Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. DNA Research. 2015; 22(6):387–401. https://doi.org/10.1093/dnares/dsv021 PMID: 26428853
7. Rosic S, Kohler F, Erhardt S. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division (vol 207, pg 335, 2014). J Cell Biol. 2014; 207(5):673-. https://doi.org/10.1083/jcb.201404097 PMID: 25365994
8. Kursel LE, Malik HS. The cellular mechanisms and consequences of centromere drive. Curr Opin Cell Biol. 2018; 52:58–65. https://doi.org/10.1016/j.ceb.2018.01.011 PMID: 29454259
9. Bracewell R, Chatla K, Nalley MJ, Bachtrog D. Dynamic turnover of centromeres drives karyotype evolution in *Drosophila*. BioRxiv [PrePrint]. 2019:733527. [posted 2019 Aug 27] https://www.biorxiv.org/content/10.1101/733527v1.full. https://doi.org/10.1101/733527 PMID: 31524597
10. Kuhn GCS, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Research. 2008; 16(2):307–24. https://doi.org/10.1007/s10577-007-1195-1 PMID: 18266060
11. Plohl M, Meštrović N, Mravinac B. Satellite DNA evolution. Repetitive DNA: Karger Publishers; 2012. p. 126–52. https://doi.org/10.1159/000337122
12. Garrido-Ramos MA. Satellite DNA: an evolving topic. Genes-Basel. 2017; 8(9). https://doi.org/10.3390/genes8090230 PMID: 28926993

13. Ferree PM, Barbash DA. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. Plos Biol. 2009; 7(10). https://doi.org/10.1371/journal.pbio.1000234 PMID: 19859525

14. Strachan T, Webb D, Dover GA. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. Embo J. 1985; 4(7):1701–8. https://doi.org/10.1002/j.1460-2075.1985.tb03839.x PMID: 16453627

15. Bachmann L, Sperlich D. Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. Mol Biol Evol. 1993; 10(3):647–59. https://doi.org/10.1093/oxfordjournals.molbev.a040029 PMID: 8336547

16. Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GCS. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. Genome Biol Evol. 2014; 6(6):1302–13. https://doi.org/10.1093/gbe/evu108 PMID: 24858539

17. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome research. 2017; 27(5):709–21. https://doi.org/10.1101/gr.213512.116 PMID: 28373483

18. Wei KHC, Lower SE, Caldas IV, Sless TJS, Barbash DA, Clark AG. Variable rates of simple satellite gains across the *Drosophila* phylogeny. Mol Biol Evol. 2018; 35(4):925–41. https://doi.org/10.1093/molbev/msy005 PMID: 29361128

19. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012; 13(1):36. https://doi.org/10.1038/nrg3117 PMID: 22124482

20. Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics. 2013; 29(6):792–3. https://doi.org/10.1093/bioinformatics/btt054 PMID: 23376349

21. Novak P, Robledillo LA, Koblizkova A, Vrbova I, Neumann P, Macas J. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic acids research. 2017; 45(12). https://doi.org/10.1093/nar/gkx257 PMID: 28402514

22. Ruiz-Ruano FJ, Lopez-Leon MD, Cabrero J, Camacho JPM. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Scientific reports. 2016; 6. https://doi.org/10.1038/srep28333 PMID: 27385065

23. de Lima LG, Svartman M, Kuhn GCS. Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. G3-Genes Genom Genet. 2017; 7(8):2831–43. https://doi.org/10.1534/g3.117.042093 PMID: 28659292

24. Palacios-Gimenez OM, Dias GB, de Lima LG, Kuhn GCES, Ramos E, Martins C, et al. High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*. Scientific reports. 2017; 7. https://doi.org/10.1038/s41598-017-06822-8 PMID: 28743997

25. Utsunomia R, Silva DMZD, Ruiz-Ruano FJ, Goes CAG, Melo S, Ramos LPE, et al. Satellitome landscape analysis of *Megaleporinus macrocephalus* (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. Scientific reports. 2019; 9. https://doi.org/10.1038/s41598-019-42383-8 PMID: 30971780

26. Liu Q, Li XY, Zhou XY, Li MZ, Zhang FJ, Schwarzacher T, et al. The repetitive DNA landscape in *Avena* (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. Bmc Plant Biol. 2019; 19. https://doi.org/10.1186/s12870-019-1769-z PMID: 31146681

27. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of *Drosophila melanogaster*. Science. 2000; 287(5461):2185–95. Epub 2000/03/25. https://doi.org/10.1126/science.287.5461.2185 PMID: 10731132

28. Bosco G, Campbell P, Leiva-Neto JT, Markow TA. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics. 2007; 177(3):1277–90. https://doi.org/10.1534/genetics.107.075069 PMID: 18039867

29. Miklos G. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. Molecular evolutionary genetics. 1985:241–321.

30. Gregory TR, Johnston JS. Genome size diversity in the family *Drosophilidae*. Heredity. 2008; 101(3):228–38. https://doi.org/10.1038/hdy.2008.49 PMID: 18523443

31. Craddock EM, Gall JG, Jonas M. Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. Genetica. 2016; 144(1):107–24. Epub 2016/01/23. https://doi.org/10.1007/s10709-016-9882-5 PMID: 26790663

32. Gall JG, Cohen EH, Polan ML. Repetitive DNA sequences in *Drosophila*. Chromosoma. 1971; 33(3):319-+. https://doi.org/10.1007/BF00284948 PMID: 5088497

33. Gall JG, Atherton DD. Satellite DNA sequences in *Drosophila virilis*. J Mol Biol. 1974; 85(4):633–64. https://doi.org/10.1016/0022-2836(74)90321-0 PMID: 4854195

34. Heikkinen E, Launonen V, Muller E, Bachmann L. The pvB370 BamHI satellite DNA family of the *Drosophila virilis* group and its evolutionary relation to mobile dispersed genetic pDv elements. Journal of molecular evolution. 1995; 41(5):604–14. https://doi.org/10.1007/BF00175819 PMID: 7490775

35. Biessmann H, Zurovcova M, Yao JG, Lozovskaya E, Walter MF. A telomeric satellite in *Drosophila virilis* and its sibling species. Chromosoma. 2000; 109(6):372–80. https://doi.org/10.1007/s004120000094 PMID: 11072792

36. Dias GB, Heringer P, Svartman M, Kuhn GC. Helitrons shaping the genomic architecture of *Drosophila*: enrichment of DINE-TR1 in alpha and beta-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology. 2015; 23(3):597–613. Epub 2015/09/27. https://doi.org/10.1007/s10577-015-9480-x PMID: 26408292

37. Abdurashitov MA, Gonchar DA, Chernukhin VA, Tomilov VN, Tomilova JE, Schostak NG, et al. Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC genomics. 2013; 14:771. Epub 2013/11/12. https://doi.org/10.1186/1471-2164-14-771 PMID: 24209985

38. Garcia G, Rios N, Gutierrez V. Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus *Austrolebias* (Cyprinodontiformes, Rivulidae). Genetica. 2015; 143(3):353–60. Epub 2015/03/21. https://doi.org/10.1007/s10709-015-9834-5 PMID: 25792372

39. Robledillo LÁ, Koblížková A, Novák P, Böttinger K, Vrbová I, Neumann P, et al. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. Scientific reports. 2018; 8(1):5838. https://doi.org/10.1038/s41598-018-24196-3 PMID: 29643436

40. Ugarkovic D, Plohl M. Variation in satellite DNA profiles—causes and effects. Embo J. 2002; 21 (22):5955–9. https://doi.org/10.1093/emboj/cdf612 PMID: 12426367

41. Morales-Hojas R, Reis M, Vieira CP, Vieira J. Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Mol Phylogenet Evol. 2011; 60(2):249–58. https://doi.org/10.1016/j.ympev.2011.04.022 PMID: 21571080

42. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research. 2016; 44(W1):W3–W10. https://doi.org/10.1093/nar/gkw343 PMID: 27137889

43. Fonseca NA, Morales-Hojas R, Reis M, Rocha H, Vieira CP, Nolte V, et al. *Drosophila americana* as a model species for comparative studies on the molecular basis of phenotypic variation. Genome Biol Evol. 2013; 5(4):661–79. https://doi.org/10.1093/gbe/evt037 PMID: 23493635

44. Baimai V. Chromosomal Polymorphisms of Constitutive Heterochromatin and inversions in *Drosophila*. Genetics. 1977; 85(1):85–93. PMID: 838273

45. Ashburner M. Drosophila. A laboratory handbook: Cold spring harbor laboratory press; 1989. ISBN: 0879693215

46. Flynn JM, Long M, Wing RA, Clark AG. Evolutionary dynamics of abundant 7 bp satellites in the genome of *Drosophila virilis*. BioRxiv [PrePrint]. 2019:693077 [posted 2019 July 4] https://www.biorxiv.org/content/10.1101/693077v1.full. https://doi.org/10.1101/693077

47. Gubenko IS, Evgenev MB. Cytological and linkage maps of *Drosophila virilis* chromosomes. Genetica. 1984; 65(2):127–39. https://doi.org/10.1007/BF00135277.

48. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome biology. 2013; 14(1). https://doi.org/10.1186/gb-2013-14-1-r10 PMID: 23363705

49. Zelentsova ES, Vashakidze RP, Krayev AS, Evgenev MB. Dispersed repeats in *Drosophila virilis*: elements mobilized by interspecific hybridization. Chromosoma. 1986; 93(6):469–76. https://doi.org/10.1007/BF00386786

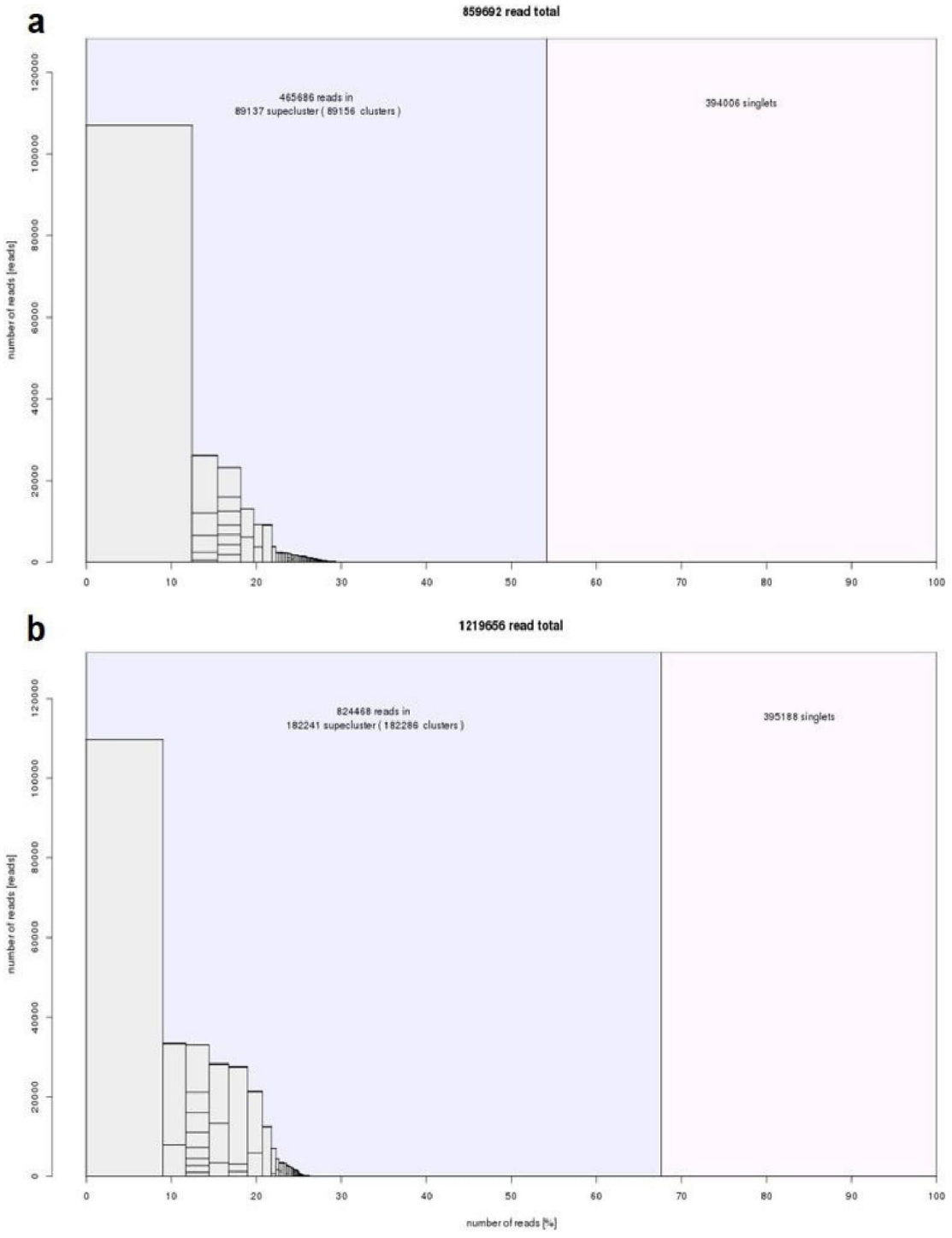
50. Roy V, Monti-Dedieu L, Chaminade N, Siljak-Yakovlev S, Aulard S, Lemeunier F, et al. Evolution of the chromosomal location of rDNA genes in two *Drosophila* species subgroups: *ananassae* and *melanogaster*. Heredity. 2005; 94(4):388. https://doi.org/10.1038/sj.hdy.6800612 PMID: 15726113

51. Wei KH, Grenier JK, Barbash DA, Clark AG. Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2014; 111(52):18793–8. https://doi.org/10.1073/pnas.1421951112 PMID: 25512552

52. Cohen EH, Bowman SC. Detection and location of three simple sequence DNAs in polytene chromosomes from *virilis* group species of *Drosophila*. Chromosoma. 1979; 73(3):327–55. Epub 1979/08/01. https://doi.org/10.1007/BF00288696 PMID: 510073

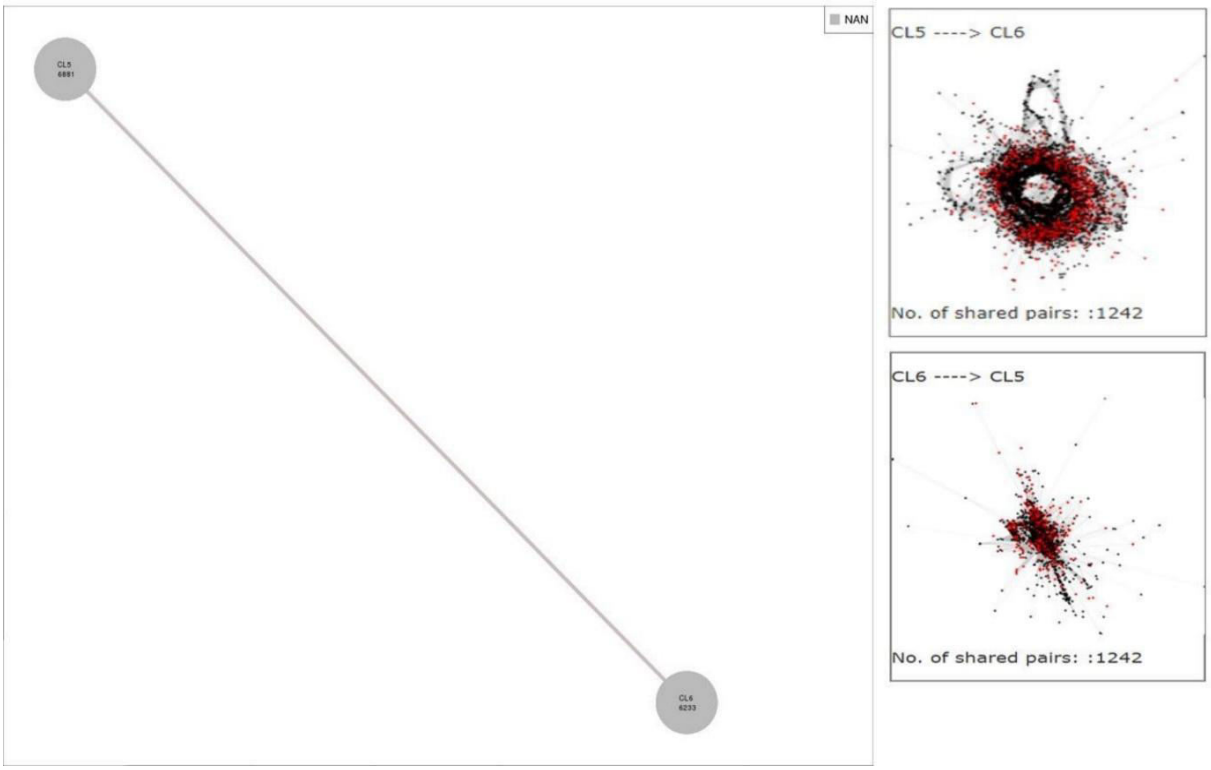
53. Mestrovic N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Satovic E, Plohl M. Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Research. 2015; 23(3):583–96. https://doi.org/10.1007/s10577-015-9483-7 PMID: 26293606

S1 Fig. TAREAN histogram summary analyses of (A) *Drosophila virilis* (strain 160) and (B) *Drosophila americana* (strain H5). The histogram analysis is the overall result of the clustering process, after filtering and pre-processing of raw reads. It shows (on the top), the total number of reads analyzed during the run. Each column represents a cluster (by abundance from left to right). The y-axis refers to the number of reads by cluster and the x-axis the percentage of each cluster in the analysis. https://doi.org/10.1371/journal.pone.0223466.s001



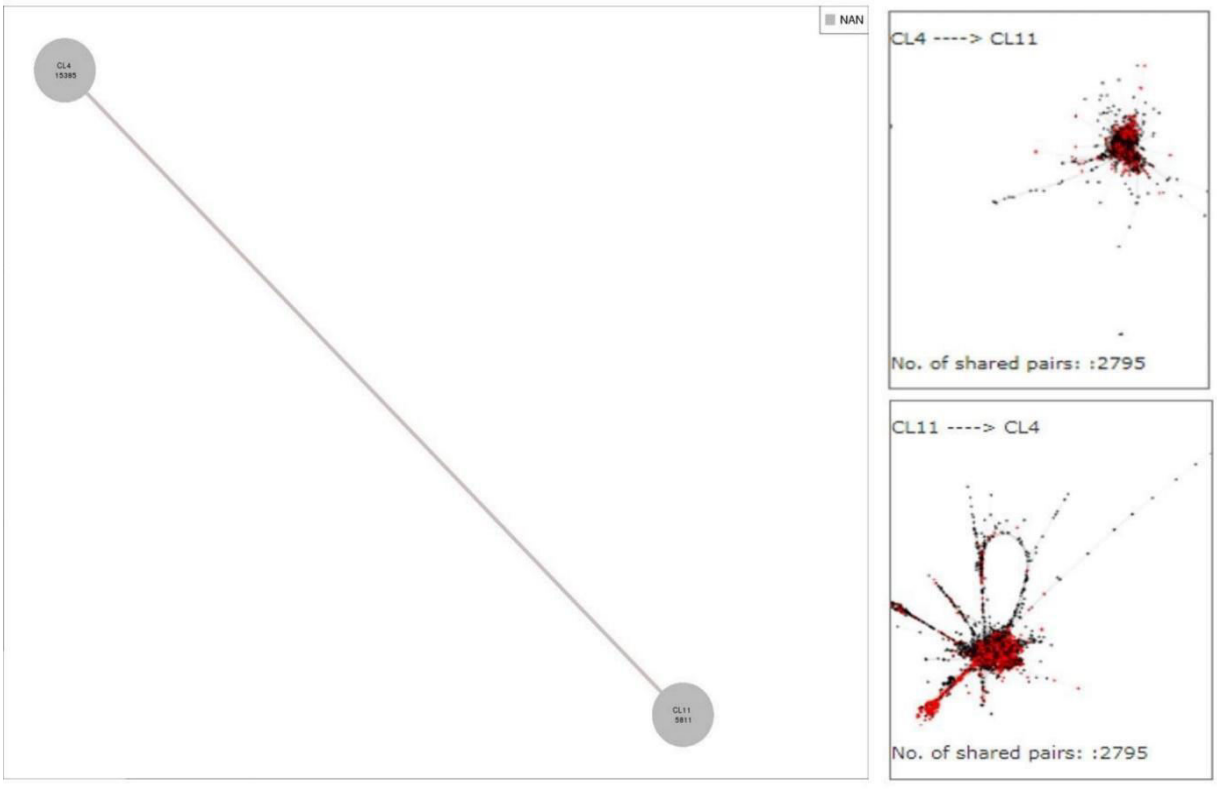
S2 Fig. pvB370 and 36TR supercluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s002



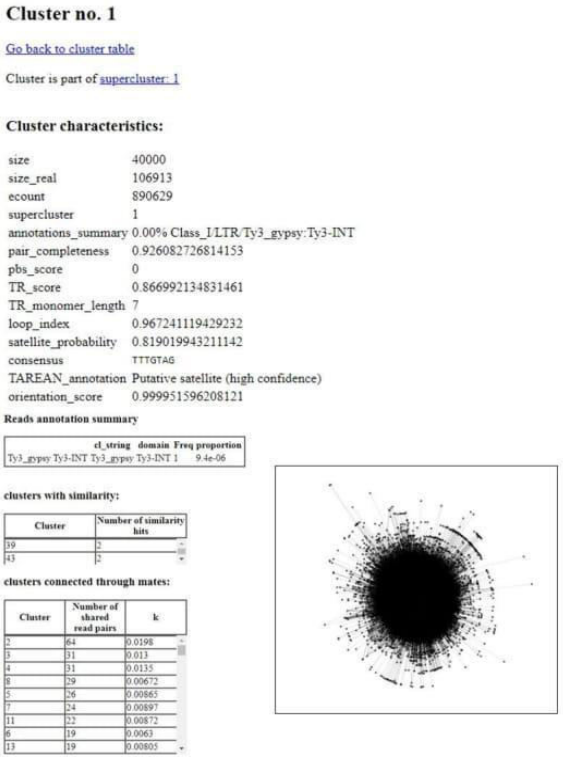
S3 Fig. pvB370 and 36TR supercluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s003



S4 Fig. Sat1 cluster analysis in *Drosophila virilis* strain 160

https://doi.org/10.1371/journal.pone.0223466.s004



Cluster no. 1

Go back to cluster table

Cluster is part of supercluster: 1

Cluster characteristics:

size	40000
size_real	106913
ecount	890629
supercluster	1
annotations_summary	0.00% Class_I/LTR/Ty3_gypsy:Ty3-INT
pair_completeness	0.926082726814153
pbs_score	0
TR_score	0.866992134831461
TR_monomer_length	7
loop_index	0.967241119429232
satellite_probability	0.819019943211142
consensus	TTTGTAG
TAREAN_annotation	Putative satellite (high confidence)
orientation_score	0.999951596208121

Reads annotation summary

	cl_string	domain	Freq proportion
Ty3_gypsy Ty3-INT	Ty3_gypsy Ty3-INT	1	9.4e-06

clusters with similarity:

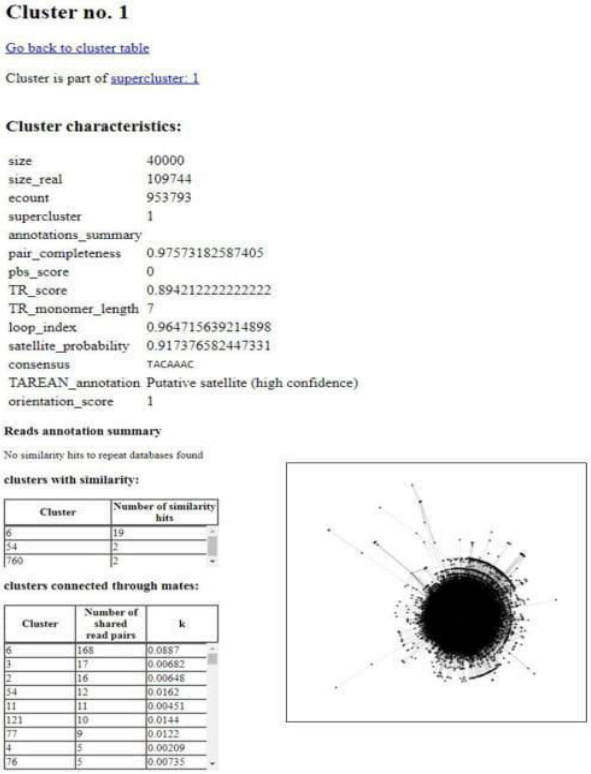
Cluster	Number of similarity hits
39	2
43	2

clusters connected through mates:

Cluster	Number of shared read pairs	k
2	64	0.0198
3	31	0.013
4	31	0.0135
8	29	0.00672
5	26	0.00865
7	24	0.00897
11	22	0.00872
6	19	0.0063
13	19	0.00805

S5 Fig. Sat1 cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s005



Cluster no. 1

Go back to cluster table

Cluster is part of supercluster: 1

Cluster characteristics:

size	40000
size_real	109744
ecount	953793
supercluster	1
annotations_summary	
pair_completeness	0.97573182587405
pbs_score	0
TR_score	0.894212222222222
TR_monomer_length	7
loop_index	0.964715639214898
satellite_probability	0.917376582447331
consensus	TACAAAC
TAREAN_annotation	Putative satellite (high confidence)
orientation_score	1

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

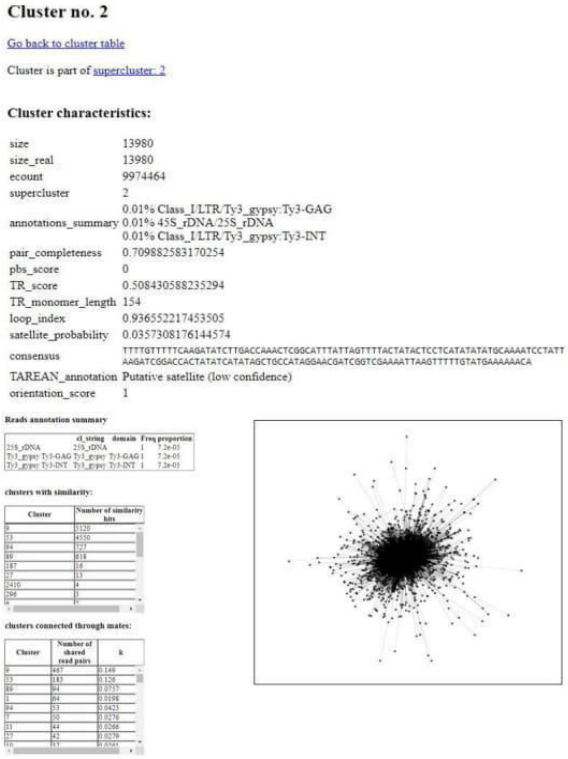
Cluster	Number of similarity hits
6	19
54	2
760	2

clusters connected through mates:

Cluster	Number of shared read pairs	k
6	168	0.0887
3	17	0.00682
2	16	0.00648
54	12	0.0162
11	11	0.00451
121	10	0.0144
77	9	0.0122
4	5	0.00209
76	5	0.00735

S6 Fig. 154TR cluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s006



Cluster no. 2

Go back to cluster table

Cluster is part of supercluster: 2

Cluster characteristics:

size	13980
size_real	13980
ecount	9974464
supercluster	2
annotations_summary	0.01% Class_I/LTR/Ty3_gypsy:Ty3-GAG 0.01% 45S_rDNA/25S_rDNA 0.01% Class_I/LTR/Ty3_gypsy:Ty3-INT
pair_completeness	0.709882583170254
pbs_score	0
TR_score	0.508430588235294
TR_monomer_length	154
loop_index	0.936552217453505
satellite_probability	0.0357308176144574
consensus	TTTTGTTTTTCAAGATATCTTGACCAAACTCGGCATTTATTAGTTTTACTATACTCCTCATATATATGCAAAATCCTATT AAGATCGGACCACTATATCATATAGCTGCCATAGGAACGATCGGTCGAAAATTAAGTTTTTGTATGAAAAAACA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	1

Reads annotation summary

	cl_string	domain	Freq	proportion
25S_rDNA	25S_rDNA		1	7.2e-05
Ty3_gypsy:Ty3-GAG	Ty3_gypsy	Ty3-GAG	1	7.2e-05
Ty3_gypsy:Ty3-INT	Ty3_gypsy	Ty3-INT	1	7.2e-05

clusters with similarity:

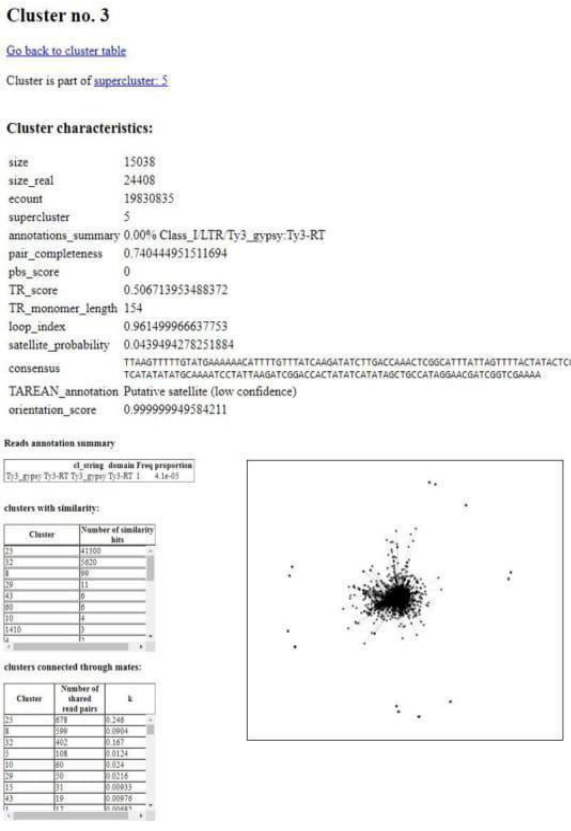
Cluster	Number of similarity hits
0	5120
53	4150
94	727
89	618
187	18
37	13
2410	4
266	1

clusters connected through mates:

Cluster	Number of shared read pairs	k
0	447	0.149
53	183	0.126
89	84	0.0737
1	64	0.0198
94	53	0.0423
7	50	0.0276
53	44	0.0266
127	42	0.0279
116	33	0.0343

S7 Fig. 154TR cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s007



Cluster no. 3

Go back to cluster table

Cluster is part of supercluster: 5

Cluster characteristics:

size	15038
size_real	24408
ecount	19830835
supercluster	5
annotations_summary	0.00% Class_I/LTR/Ty3_gypsy:Ty3-RT
pair_completeness	0.740444951511694
pbs_score	0
TR_score	0.506713953488372
TR_monomer_length	154
loop_index	0.961499966637753
satellite_probability	0.0439494278251884
consensus	TTAAGTTTTTGTATGAAAAAACATTTTGTTTATCAAGATATCTTGACCAAACTCGGCATTTATTAGTTTTACTATACTCC TCATATATATGCAAAATCCTATTAAGATCGGACCACTATATCATATAGCTGCCATAGGAACGATCGGTCGAAAA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	0.999999949584211

Reads annotation summary

	cl_string	domain	Freq	proportion
Ty3_gypsy:Ty3-RT	Ty3_gypsy	Ty3-RT	1	4.1e-05

clusters with similarity:

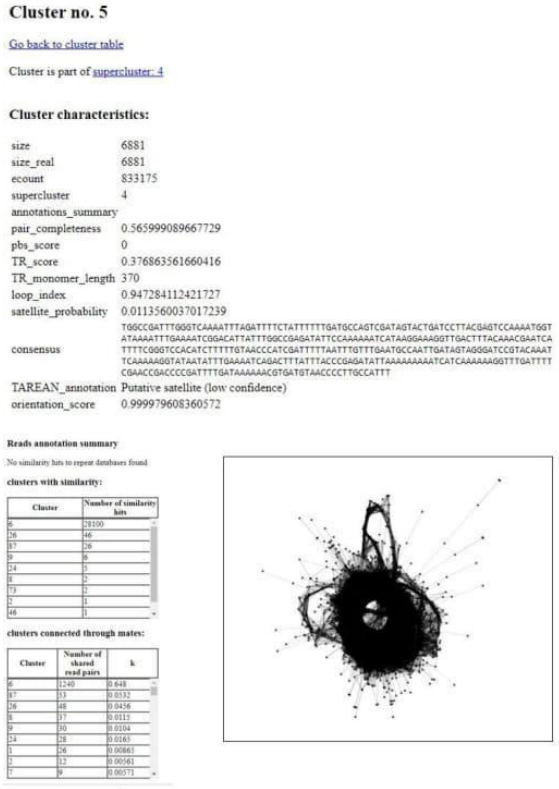
Cluster	Number of similarity hits
23	41300
32	5620
8	99
29	11
43	8
60	6
10	4
1410	3
4	1

clusters connected through mates:

Cluster	Number of shared read pairs	k
23	678	0.248
8	599	0.0904
32	402	0.167
5	108	0.0124
10	60	0.024
29	30	0.0216
15	31	0.00933
43	19	0.00976
6	17	0.00487

S8 Fig. pvB370 cluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s008



Cluster no. 5

Go back to cluster table

Cluster is part of supercluster: 4

Cluster characteristics:

size	6881
size_real	6881
ecount	833175
supercluster	4
annotations_summary	
pair_completeness	0.565999089667729
pbs_score	0
TR_score	0.376863561660416
TR_monomer_length	370
loop_index	0.947284112421727
satellite_probability	0.0113560037017239
consensus	TGGCCGATTTGGGTCAAAATTTAGATTTTCTATTTTTTGATGCCAGTCGATAGTACTGATCCTTACGAGTCCAAAATGGT ATAAAATTTGAAAATCGGACATTATTTGGCCGAGATATTCCAAAAAATCATAAGGAAAGGTTGACTTTACAAACGAATCA TTTTCGGGTCCACATCTTTTTGTAACCCATCGATTTTTAATTTGTTTGAATGCCAATTGATAGTAGGGATCCGTACAAAT TCAAAAAGGTATAATATTTGAAAATCAGACTTTATTTACCCGAGATATTAAAAAAAATCATCAAAAAAGGTTTGATTTT CGAACCGACCCCGATTTTGATAAAAAACGTGATGTAACCCCTTGCCATTT
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	0.999979608360572

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

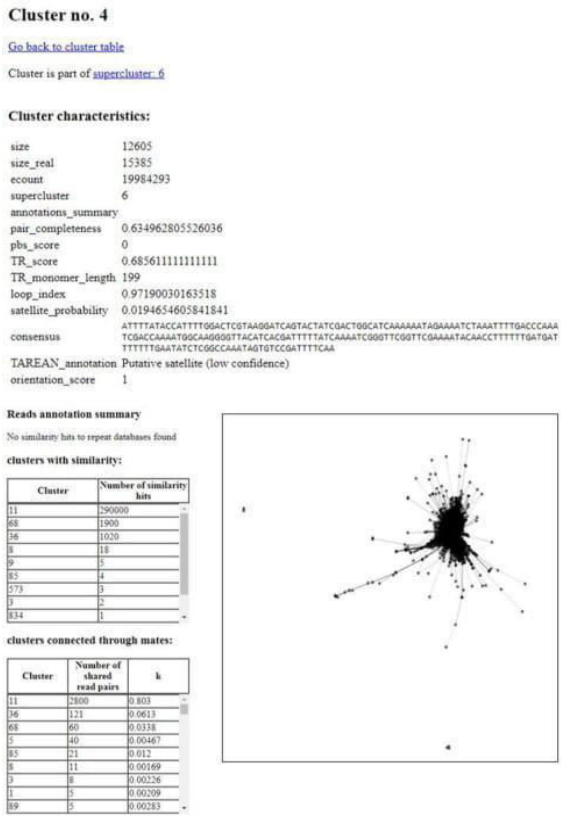
Cluster	Number of similarity hits
6	28100
28	46
87	26
9	6
24	5
8	2
73	2
2	1
46	1

clusters connected through mates:

Cluster	Number of shared read pairs	k
6	1240	0.648
87	53	0.0532
28	48	0.0456
8	17	0.0115
9	30	0.0104
24	28	0.0165
1	26	0.00863
2	12	0.00561
7	9	0.00573

S9 Fig. pvB370 cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s009



Cluster no. 4

Go back to cluster table

Cluster is part of supercluster: 6

Cluster characteristics:

size	12605
size_real	15385
ecount	19984293
supercluster	6
annotations_summary	
pair_completeness	0.634962805526036
pbs_score	0
TR_score	0.685611111111111
TR_monomer_length	199
loop_index	0.97190030163518
satellite_probability	0.0194654605841841
consensus	ATTTTATACCATTTTGGACTCGTAAGGATCAGTACTATCGACTGGCATCAAAAAATAGAAAATCTAAATTTTGACCCAAA TCGACCAAAATGGCAAGGGGTTACATCACGATTTTTATCAAAATCGGGTTCGGTTCGAAAATACAACCTTTTTGATGAT TTTTTTGAATATCTCGGCCAAATAGTGTCCGATTTTCAA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	1

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

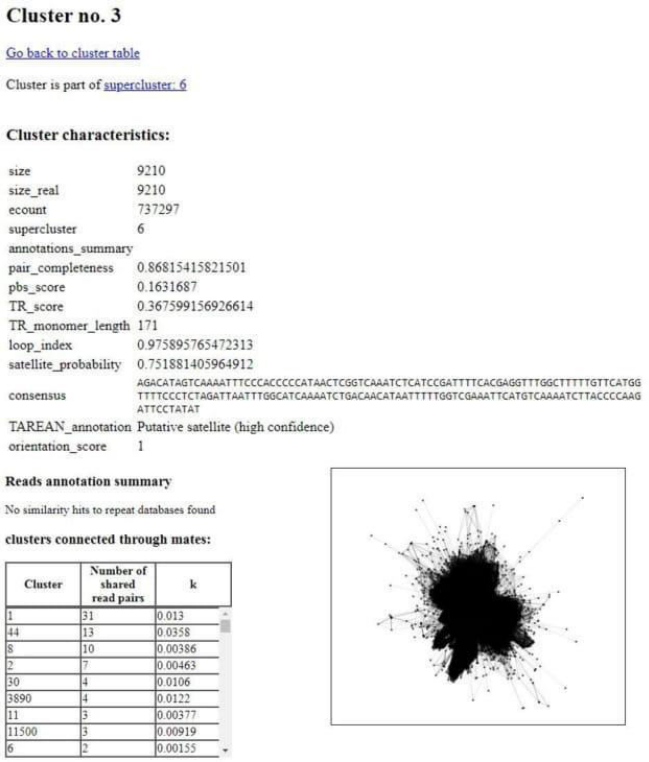
Cluster	Number of similarity hits
11	290000
68	1900
36	1020
5	18
9	5
85	4
573	3
3	2
834	1

clusters connected through mates:

Cluster	Number of shared read pairs	k
11	2800	0.803
36	121	0.0613
68	60	0.0338
5	40	0.00467
85	21	0.012
8	11	0.00169
3	8	0.00226
1	5	0.00209
89	5	0.00283

S10 Fig. 172TR cluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s010



Cluster no. 3

Go back to cluster table

Cluster is part of supercluster: 6

Cluster characteristics:

size	9210
size_real	9210
ecount	737297
supercluster	6
annotations_summary	
pair_completeness	0.86815415821501
pbs_score	0.1631687
TR_score	0.367599156926614
TR_monomer_length	171
loop_index	0.975895765472313
satellite_probability	0.751881405964912
consensus	AGACATAGTCAAAATTTCCCACCCCCATAACTCGGTCAAATCTCATCCGATTTTCACGAGGTTTGGCTTTTTGTTCATGG TTTTCCCTCTAGATTAATTTGGCATCAAAATCTGACAACATAATTTTTGGTCGAAATTCATGTCAAAATCTTACCCCAAG ATTCCTATAT
TAREAN_annotation	Putative satellite (high confidence)
orientation_score	1

Reads annotation summary

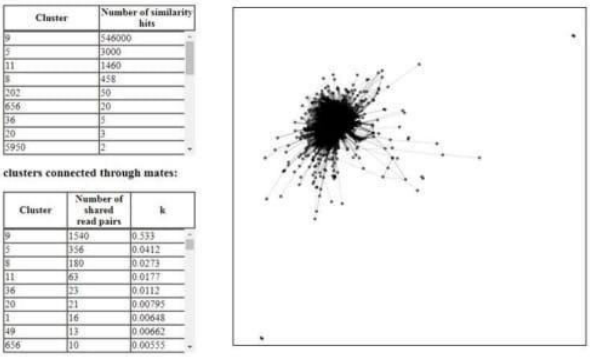
No similarity hits to repeat databases found

clusters connected through mates:

Cluster	Number of shared read pairs	k
1	31	0.013
44	13	0.0358
8	10	0.00386
2	7	0.00463
30	4	0.0106
3890	4	0.0122
11	3	0.00377
11500	3	0.00919
6	2	0.00155

S11 Fig. 172TR cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s011



Cluster no. 2

Go back to cluster table

Cluster is part of supercluster: 2

Cluster characteristics:

size	13646
size_real	25372
ecount	20004210
supercluster	2
annotations_summary	
pair_completeness	0.752209944751381
pbs_score	0
TR_score	0.58228
TR_monomer_length	171
loop_index	0.967457577315801
satellite_probability	0.0488727035644302
consensus	TAACTCGGTCAAATCTCATCCGATTTTCACGAAGTTTGGCTTTTTCTTCATGGTTTTCCCTCTAGATTAATCTGGCATTC AAATCTCGACACCGATTTTTTAGTCGAAATTCATGTCCAAATCTTACCCCAAGATTCGTATATAGACATGGGTCAAAAAT TACCACCCCCA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	1

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

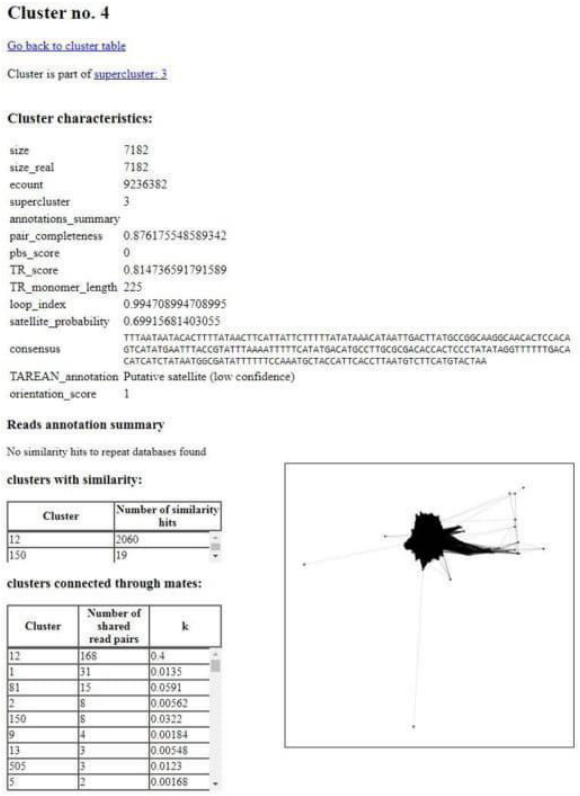
Cluster	Number of similarity hits
9	546000
5	3000
11	1460
8	458
202	50
656	20
36	5
20	3
5950	2

clusters connected through mates:

Cluster	Number of shared read pairs	k
9	1540	0.533
5	356	0.0412
8	180	0.0273
11	63	0.0177
36	23	0.0112
20	21	0.00795
1	16	0.00648
49	13	0.00662
656	10	0.00555

S12 Fig. 225TR cluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s012



Cluster no. 4

Go back to cluster table

Cluster is part of supercluster: 3

Cluster characteristics:

size	7182
size_real	7182
ecount	9236382
supercluster	3
annotations_summary	
pair_completeness	0.876175548589342
pbs_score	0
TR_score	0.814736591791589
TR_monomer_length	225
loop_index	0.994708994708995
satellite_probability	0.69915681403055
consensus	TTTAATAATACACTTTTATAACTTCATTATTCTTTTTATATAAACATAATTGACTTATGCCGGCAAGGCAACACTCCACA GTCATATGAATTTACCGTATTTAAAATTTTTCATATGACATGCCTTGCGCGACACCACTCCCTATATAGGTTTTTTGACA CATCATCTATAATGGCGATATTTTTTTCCAAATGCTACCATTCACCTTAATGTCTTCATGTACTAA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	1

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

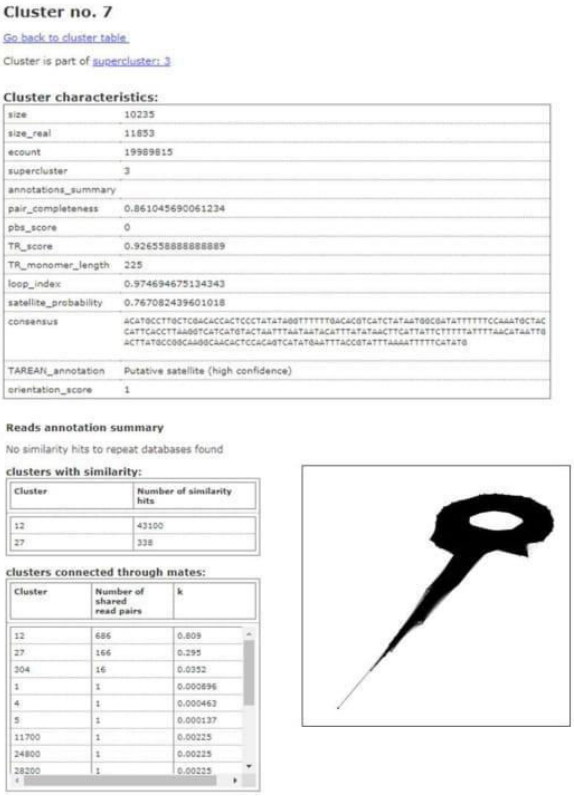
Cluster	Number of similarity hits
12	2060
150	19

clusters connected through mates:

Cluster	Number of shared read pairs	k
12	168	0.4
1	31	0.0135
81	15	0.0591
2	8	0.00562
150	8	0.0322
9	4	0.00184
13	3	0.00548
505	3	0.0123
5	2	0.00168

S13 Fig. 225TR cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s013



Cluster no. 7

Go back to cluster table

Cluster is part of supercluster: 3

Cluster characteristics:

size	10235
size_real	11853
ecount	19989815
supercluster	3
annotations_summary	
pair_completeness	0.861045690061234
pbs_score	0
TR_score	0.926558888888889
TR_monomer_length	225
loop_index	0.974694675134343
satellite_probability	0.767082439601018
consensus	ACATGCCTTGCTCGACACCACTCCCTATATAGGTTTTTTGACACGTCATCTATAATGGCGATATTTTTTCCAAATGCTAC CATTCACCTTAAGGTCATCATGTACTAATTTAATAATACATTTATATAACTTCATTATTCTTTTTATTTTAACATAATTG ACTTATGCCGGCAAGGCAACACTCCACAGTCATATGAATTTACCGTATTTAAAATTTTTCATATG
TAREAN_annotation	Putative satellite (high confidence)
orientation_score	1

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

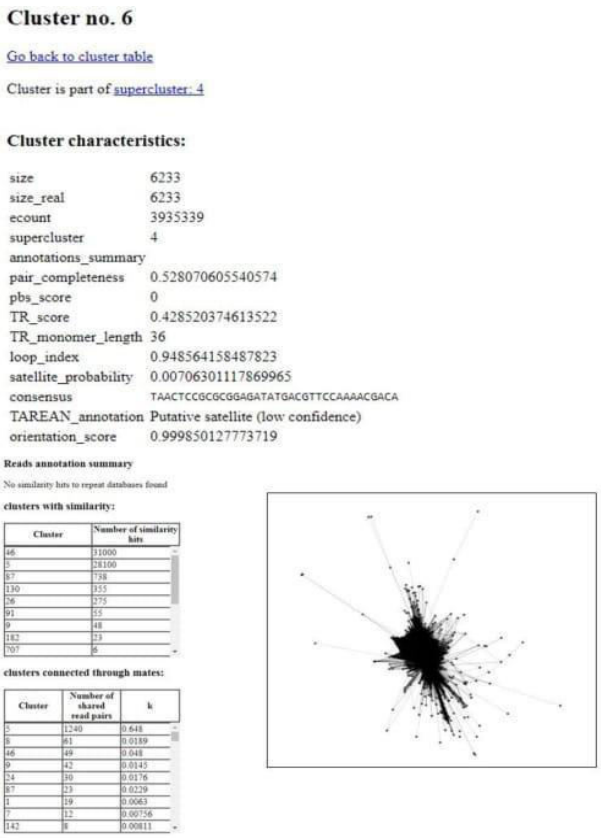
Cluster	Number of similarity hits
12	43100
27	338

clusters connected through mates:

Cluster	Number of shared read pairs	k
12	686	0.809
27	166	0.295
304	16	0.0352
1	1	0.000896
4	1	0.000463
5	1	0.000137
11700	1	0.00225
24800	1	0.00225
28200	1	0.00225

S14 Fig. 36TR cluster analysis in *Drosophila virilis* strain 160.

https://doi.org/10.1371/journal.pone.0223466.s014



Cluster no. 6

Go back to cluster table

Cluster is part of supercluster: 4

Cluster characteristics:

size	6233
size_real	6233
ecount	3935339
supercluster	4
annotations_summary	
pair_completeness	0.528070605540574
pbs_score	0
TR_score	0.428520374613522
TR_monomer_length	36
loop_index	0.948564158487823
satellite_probability	0.00706301117869965
consensus	TAACTCCGCGCGGAGATATGACGTTCCAAAACGACA
TAREAN_annotation	Putative satellite (low confidence)
orientation_score	0.999850127773719

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

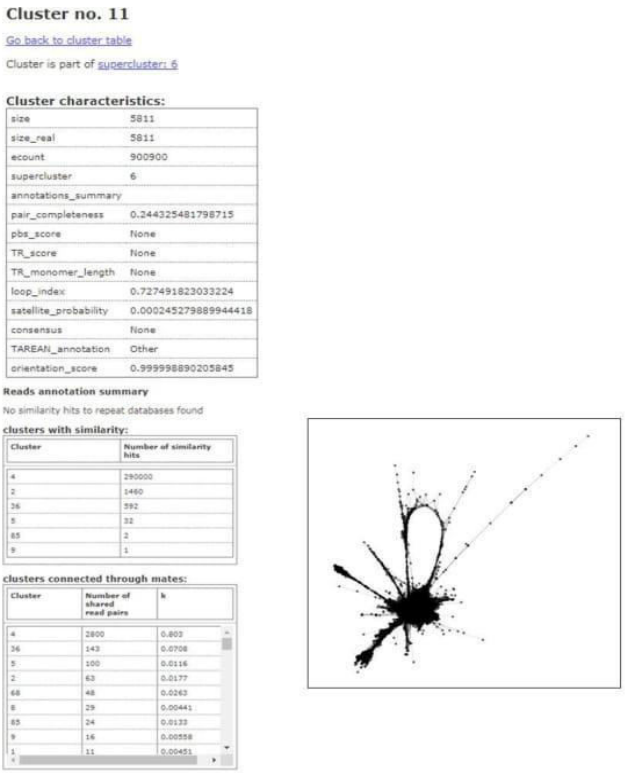
Cluster	Number of similarity hits
46	31000
5	28100
87	738
130	355
26	275
91	55
9	48
182	23
707	6

clusters connected through mates:

Cluster	Number of shared read pairs	k
5	1240	0.648
8	61	0.0189
46	49	0.048
9	42	0.0145
24	30	0.0176
87	23	0.0229
1	19	0.0063
7	12	0.00756
142	8	0.00811

S15 Fig. 36TR cluster analysis in *Drosophila americana* strain H5.

https://doi.org/10.1371/journal.pone.0223466.s015



Cluster no. 11

Go back to cluster table

Cluster is part of supercluster: 6

Cluster characteristics:

size	5811
size_real	5811
ecount	900900
supercluster	6
annotations_summary	
pair_completeness	0.244325481798715
pbs_score	None
TR_score	None
TR_monomer_length	None
loop_index	0.727491823033224
satellite_probability	0.000245279889944418
consensus	None
TAREAN_annotation	Other
orientation_score	0.999998890205845

Reads annotation summary

No similarity hits to repeat databases found

clusters with similarity:

Cluster	Number of similarity hits
4	290000
2	1460
36	592
5	32
65	2
9	1

clusters connected through mates:

Cluster	Number of shared read pairs	k
4	2800	0.803
36	143	0.0708
5	100	0.0116
2	63	0.0177
68	48	0.0263
8	29	0.00441
65	24	0.0133
9	16	0.00558
1	11	0.00451

4. Capítulo 2: Manuscript

"Identification and comparative analyses of satellite DNAs offers new insights for phylogenetics hypotheses between *Drosophila* species from the *montium* group"

Bráulio S. M. L. Silva, Marta Svartman and Gustavo C. S. Kuhn
Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil
Corresponding author: gcskuhn@ufmg.br

Abstract

Satellite DNAs (satDNAs) are abundant repetitive sequences organized as long arrays of tandem repeats which may display important functional roles in eukaryotic genomes, such as gene regulation, chromatin modulation and establishment of functional centromeres. Before the genomic era, studies involving the identification of satDNAs involved biased experimental approaches that were inefficient to retrieve all satDNAs from a given genome. Currently, the genome sequencing by Next-generation sequencing (NGS) techniques and the development of new softwares and pipelines has enabled the identification of all satDNAs (satellitome) from a given species. In the genus *Drosophila*, satDNAs sequences are well studied and characterized in some groups of species, such as in the *melanogaster* group, *virilis* group and *repleta* group, for example. In the *montium* group, one of the largest *Drosophila* groups, presenting more than 90 Asian and Australasian species, satDNA studies are virtually inexistent. The phylogeny between species in the *montium* group is still under investigation but recent studies shed a new light on this topic. Here, we used the Tandem Repeat Analyzer (TAREAN) pipeline for a *de novo* identification of satDNAs in 23 recent sequenced species from the *montium* group. This is the first study aiming the identification and characterization of the satellitome in species from the *montium* group. We identified 142 satDNA clusters and 17 satDNA families (TRs) shared by at least two species. These TR families are composed of different monomer sequences, which varies from 4 bp to >1000 bp and present from 0.015% to 16.0% genome proportions. Also, our results show that these genomes are enriched by simple (≤10 bp) satDNA sequences and that there exists a strong negative correlation between the monomer size and the copy number of satDNAs. Additionally, we found that our phylogenetic trees based on satDNA copies of each satDNA family are in accordance with the most up-to-date phylogenetic hypotheses proposed for species in the *montium* group and that 11 satDNA families are useful taxonomic and phylogenetic markers within the group. Lastly, we also found that some satDNA families are related to Helitron transposable elements, specially the DINE-TR1, whose internal tandem repeats gave rise to satDNA arrays in species from the *serrata* subgroup.

Introduction

Eukaryote genomes are enriched by a great number and variety of non-coding repetitive DNAs. The total abundance of these sequences varies between species, but it can reach >50% of some genomes, including humans (de Koning *et al*., 2011; López-Flores and Garrido-Ramos, 2012). Non-coding repetitive elements can be exemplified by scattered sequences, such as transposable elements (TEs), or by sequences organized in tandem, such as micro-satellites, mini-satellites and satellite DNAs (satDNAs). These different classes of tandem repeats differ from each other mainly by a combination of the repeat size, arrays size and genomic location. According to Tautz (1993), microsatellites have monomers between 1-6 bp, can be found anywhere in the genome and are repeated from 5 to 100 copies in each array. Mini-satellites can vary between 9-100 bp, are present in different regions of the genomes (mostly found in telomeric regions) and repeated hundreds of times at each locus. Meanwhile, satDNAs are generally found in heterochromatin, mainly in the centromeric regions, where they are arranged in arrays which can reach up to Megabases (Mbs). Additionally, satDNAs vary in size, with monomers ranging from 2 bp to more than 1,000 bp (>1kb).

SatDNAs do not encode proteins, but they can play important functional roles in the genome, such as in gene regulation, chromatin modulation and establishment of functional centromeres (Yunis and Yasmineh, 1971; Kuhn, 2015; Sullivan *et al*., 2017). Besides that, satDNAs usually have a fast evolution rate (Garrido-Ramos, 2017), and even phylogenetically close species may differ in satDNA content and DNA sequence (Melters *et al*., 2013). For this reason, satDNAs can be used as interesting markers for taxonomy or phylogenetic inferences among closely related species.

Experimental methodologies for repetitive DNA identification in the pre-genomic era were time-consuming, limited and laborious. Currently, the genomes of several species are being sequenced by massive new generation DNA sequencing techniques (Next Generation Sequencing). In the *Drosophila* genus, many species have already been sequenced and the number of available genomes is increasing fast (Adams *et al*., 2000; Garrigan *et al*., 2012; Miller *et al*., 2018; Bronski *et al*., 2020).

The *Drosophila montium* group is currently composed of 94 Asian and Australasian species (Yassin, 2018). Traditionally, mainly based on morphological characters, the group has been classified as a subgroup within the *melanogaster* group (Bock and Wheeler, 1972; Ashburner *et al*., 1983). However, a more recent study repositioned the clade as an independent group (Da Lage *et al*., 2007). According to Russo *et al*., (2013), the separation between *melanogaster* and *montium* groups took place about 27 Mya. Recently, Yassin (2018) analyzed morphological (male abdominal pigmentation and genitalia) and chorological traits and subdivided the group into seven subgroups: *parvula*, *montium*, *punjabiensis*, *serrata*, *kikkawai* and *seguyi*. Additionally, the *montium* group is divided into 8 complexes, but there are still some species and clades as *incertae sedis* (Yassin,

2018).

It is assumed that the common ancestor of all subgroups from the *montium* group lived in Asia approximately 19.3 Mya. Posteriorly, the lineage split giving rise to two clades: the *parvula* subgroup and the clade including the ancestor of all the others seven subgroups. The rising of the Himalayas was responsible for the split between the *montium* subgroup and the *punjabiensis*, *serrata*, *kikkawai* and *seguyi* subgroups (Yassin, 2018).

More recently, Conner *et al*., (2021) analyzed 60 nuclear genes and confirmed the monophyly of the seven subgroups proposed by Yassin (2018). However, differently from what was proposed by Yassin (2018), who included the *parvula* subgroup as the most basal subgroup in the phylogeny, Conner *et al*., (2021) proposed that the *montium* subgroup is the most basal subgroup, with the *punjabiensis* subgroup closer to the *seguyi* subgroup and the *kikkawai* subgroup as the third most basal clade from the group (Figure 1) (Conner *et al*., 2021).

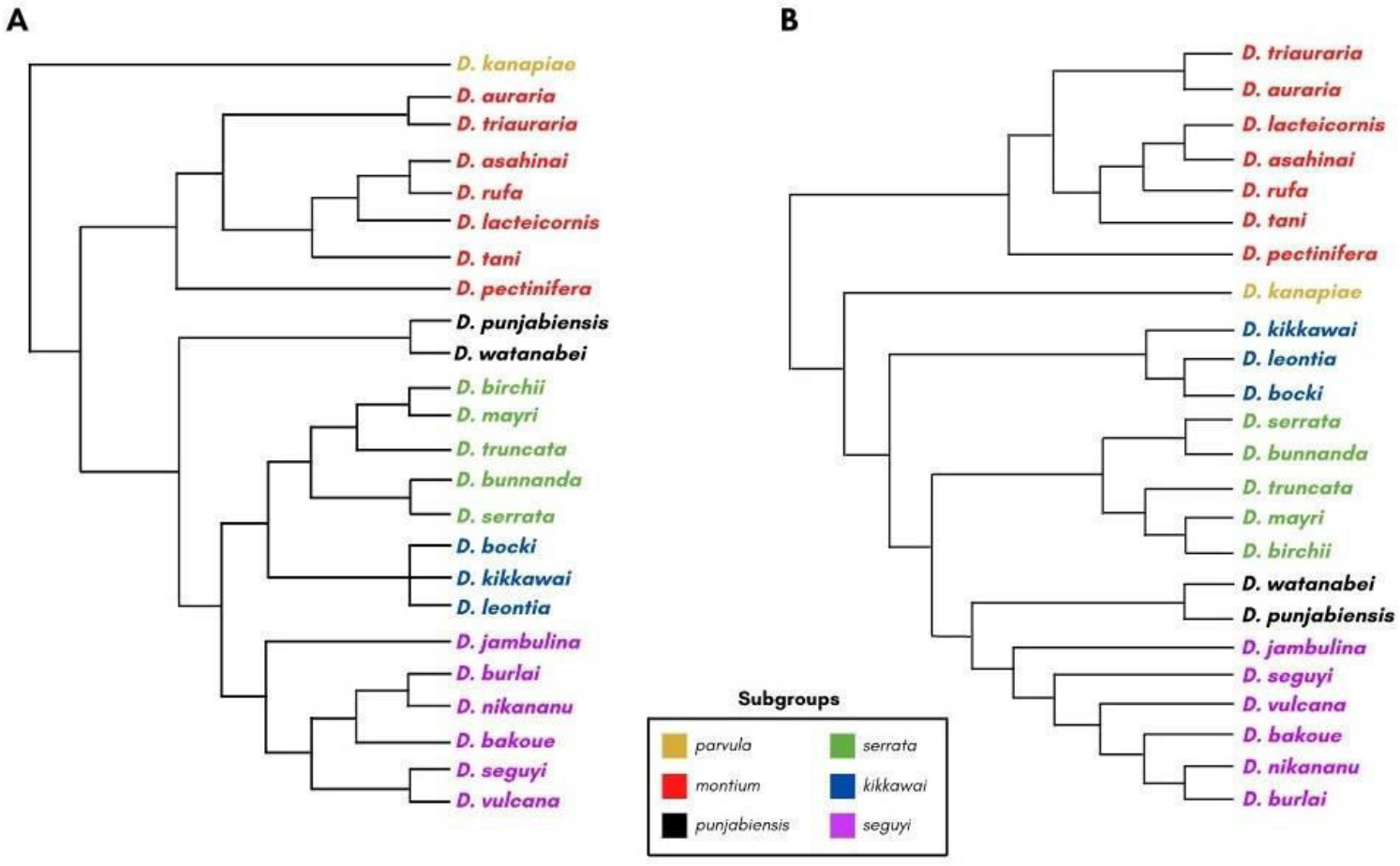


Figure 1. The two most recent phylogenetic hypotheses of the *montium* group. Only the species investigated on our work are shown. *A.* Phylogeny proposed by Yassin (2018) based on morphological and chorological data. *B*. Phylogeny proposed by Conner *et al*., (2021) based on bayesian and maximum likelihood (ML) analyses of 60 nuclear genes.

Genetic studies with species from the *montium* group have been based mainly on the comparative analysis of karyotypes. Baimai (1980), for example, found no major changes in the karyotypes of 20 species from the *montium* group, with all species showing one pair of sex chromosomes, two pairs of acrocentric chromosomes and one pair of microchromosomes. However, the author reported extensive interspecific variation in the amount of heterochromatin present in the Y chromosome and microchromosomes, and to a lesser extent in the X. Posteriorly, Venkat and Ranganath (2007) analyzed the karyotypes of four species from the *montium* group (*D. agumbensis*, *D. anomelani*, *D. truncata* and *D. cauverii*) and found the same pattern of variation restricted to the heterochromatin content of the Y chromosome and microchromosomes. According to the authors, the heterochromatic variation could be related to changes regarding the satDNA content of the species.

In recent years, the genomes of a few species from the *montium* group were sequenced by different new generation DNA sequencing approaches (Allen *et al*., 2017; Miller *et al*., 2018; Kim *et al*., 2021). Among them, the three species *D. triauraria*, *D. serrata* and *D. kikkawai*, were sequenced through Illumina, PacBio and Hi-C sequencing technologies. Most recently, using Illumina HiSeq 2000 and HiSeq 2500 Systems technologies, Bronski *et al*., (2020) and Conner *et al*., (2021) sequenced more than 40 species from the *montium* group, which allows new genomic data for comparative and evolutionary studies. The genome size estimates for these species varies from 155 Mb (*D. bocki* and *D. kanapiae*) to 223 Mb (*D. mayri*).

Different from other groups of the genus (e.g., *melanogaster, repleta* and *virilis*) there are still few studies at genomic level for the *montium* group and no comparative studies of repetitive sequences (like satellite DNAs). Thus, the availability of sequenced genomes provides important data for investigation of satDNAs and, consequently, of the genomes of these species.

With the advent of new generation DNA sequencing techniques, new bioinformatics tools have been providing efficient ways to identify and classify repetitive DNAs, including satDNAs (Dias *et al*., 2014; Dias *et al*., 2021). However, the high repeatability of these sequences constitutes a barrier during the genomic assembly process (Treangen and Salzberg, 2012). Consequently, many repetitive DNAs are not included in the assemblies, thus making it difficult their identification and characterization. One way to solve this problem is by identifying these repetitive DNAs from unmounted raw reads, generated after DNA sequencing. This approach was implemented in the computer pipeline TAREAN (Novák *et al*., 2017). TAREAN is a pipeline developed for the identification of tandemly repeated sequences from eukaryotic genomes. TAREAN has been used for the study of tandem repeats in several species of eukaryotes (García *et al*., 2015; Ruiz-Ruano *et al*., 2016; Ávila Robledillo *et al*., 2018). In a recent study in species from the *virilis* group, we also showed that TAREAN is an efficient method to detect tandem repeats, specially satDNAs from *Drosophila* genomes (Silva *et al*., 2019).

In the present work, we aimed to characterize for the first time the satDNA landscape of species from the *montium* group using TAREAN. The data is discussed in terms of satDNAs general structural features, the use of the satDNA data for phylogenetic inferences within the group and the relationship between satDNAs and Helitrons transposable elements.

Material and Methods

TAREAN analyses

TAREAN is a computational pipeline for unsupervised identification of satDNAs from unassembled short sequences reads. In this study, we used the sequencing raw data from 23 species (females) whose publicly Illumina paired-end reads were recently sequenced by the Eisen lab and published by Bronski *et al*., (2020) (Table 1). The analyses were performed on Galaxy Platform (Afgan *et al*., 2018). We measured the reads quality with the "FASTQC" tool and converted all the sequences to a single fastqsanger format with the "FASTQ Groomer" (Sanger and Illumina 1.8 +). After removal of adapters and reads presenting more than 5% of low-quality bases (Phred cutoff<10), the reads were trimmed to 100 bp long with the "Preprocessing of fastq paired-reads" tool. The resulting file with the interlaced filtered paired-end reads was used as input for the Tandem Repeat Analyzer (TAREAN) tool, with the following settings: "read sampling: no - advanced options: yes - perform cluster merging: yes - use custom repeat database: no - cluster size threshold for detailed analysis: 0.01 - perform automatic filtering of abundant satellite repeats: no - keep original read names: no - similarity search options: masking of low complexity repeats disabled - select queue: basic". Archives with the HTML reports containing the satDNA clusters were downloaded for a more detailed investigation. In this study, we only analyzed satDNA clusters making at least 0.1% of the genome in at least one species.

Satellite DNAs identification

For each sequenced genome, the TAREAN pipeline divides the resulting clusters of sequences in two categories: satellites with high confidence (HC) and satellites with low confidence (LC). These categories are determined according to the "Connected component index (C)" which indicates clusters formed by tandem repeat sequences and "Pair completeness index (P)" which measures the length of continuous tandem arrays (Novák *et al*., 2017).

The TAREAN analyses was performed for all the 23 species investigated (Table 1) and all the consensus sequences (corresponding to their respective clusters) with more than 0.1% genome proportion were analyzed. Based on the computational analyses made by Novák *et al*., (2013, 2017), we developed a new satDNA "filter" in which the clusters should comply with three out of the four following parameters: c-value>0.9, p-value>0.8, high confidence and circular graph layout. After this

cutoff analysis, we proceeded with further investigations of the remaining satDNA clusters/consensus sequences. Figure 2 shows the workflow chart of our study.

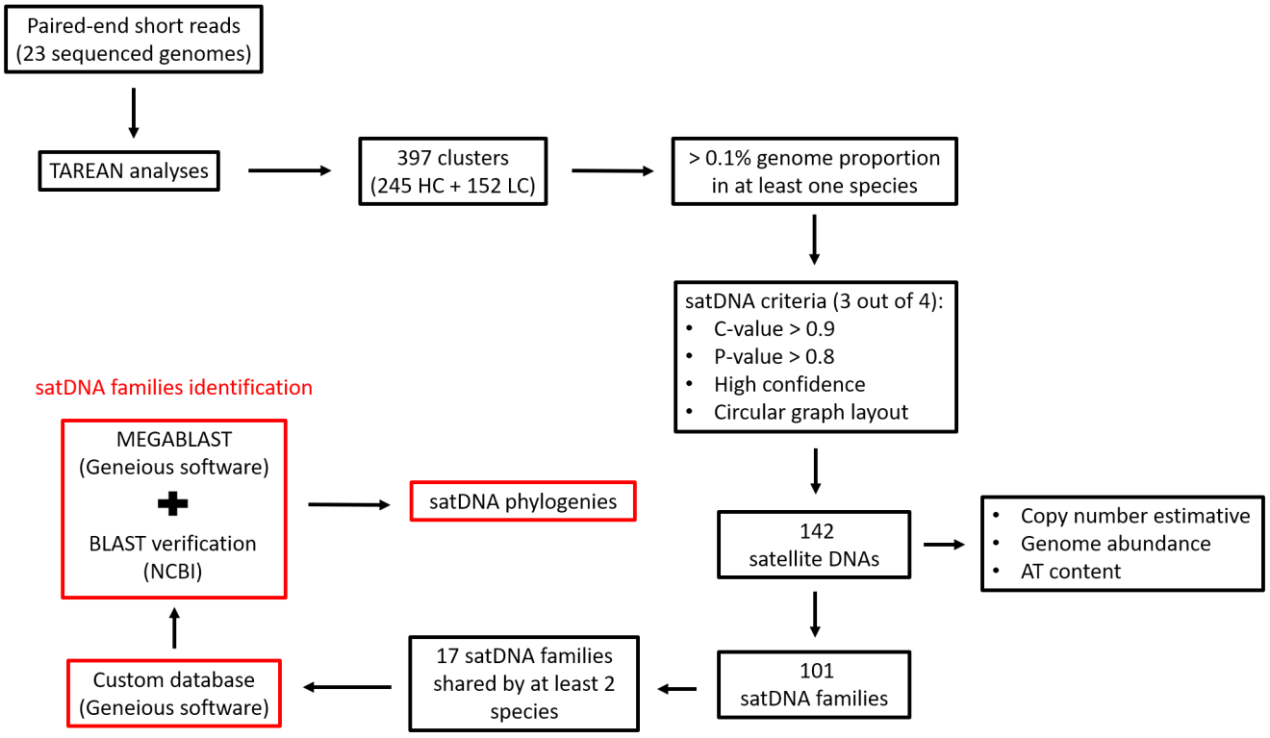


Figure 2. Workflow chart. Satellite DNA identification in the 23 sequenced *montium* genomes was divided in two main analyses: satDNAs identification (black) and satDNA families identification (red). After the TAREAN analyses and our data processing, we selected 142 satDNA sequences, including 17 that are shared by at least 2 species.

Identification of satellite DNA families

For the identification of satDNA families shared by two or more species, we created a custom database with the consensus sequences of satDNAs retrieved from the TAREAN analyses on the Geneious Software. For each consensus sequence we run MEGABLAST on the custom database (maximum e-value = 1e-5, gap cost = linear, threshold = 0%, majority: most common bases, fewest ambiguities) aiming to detect homologous sequences in the other species. BLAST searches (NCBI BLAST) using each consensus sequence as a query were also performed on the whole-genome shotgun contigs (WGS) database of all analyzed species in order to look for homologous sequences. For the BLAST, the queries started with ten tandem repeats of each consensus sequence. According to the monomer sequence length, we used different threshold values: 100% query cover + 70% identity for monomers longer than 100 bp and 100% query + 85% identity for monomers shorter than 100 bp. The identity value for short repeats (<100 bp) was determined according to previous studies

of short satDNA families present in other *Drosophila* species, e.g., *D. melanogaster* and *D. virilis* (Gall *et al*., 1971; Lohe *et al*., 1993). For each satDNA family identified, we considered that at least one cluster/consensus sequence retrieved from the TAREAN analyses must comply with the satDNA criteria defined in our work, as explained in the "satellite DNAs identification" section.

Correlation tests and phylogenies

The correlation tests were based on general structural features of the satDNAs and genome size data previously published by (Bronski *et al*., 2020). We used Spearman's correlation test to determine statistical significance. Plot graphs were developed in R language on the RStudio Software.

The phylogeny used for the heatmap on satDNA distribution and abundance was reconstructed based on Yassin (2018). The kinship relations between taxa were constructed using the "Newick three format" and the branch length was adjusted with the Archaeopteryx software (Han and Zmasek, 2009). For tandem repeat arrays examination, we run MEGABLAST experiments (WGS database) on BLAST – NCBI (Johnson *et al*., 2008) for all species studied using each satDNA consensus sequence identified. The queries consisted of 10, 100 or 500 tandem repeats and the best total score hit was downloaded for a more detailed investigation.

SatDNAs phylogenies were inferred based on the alignment of 20 satDNA repeats from each species. To better access the variability of the copies, only 5 monomers were retrieved from each hit (contig) after NCBI BLAST run. SatDNA repeats were aligned with MUSCLE method (UPGMA) (Edgar, 2004). Maximum Likelihood (ML) trees were constructed using the MEGA X Software (Tamura, 1992; Kumar *et al*., 2018) with 1000 bootstrap replicates and best-fit model predicted for each nucleotide alignment.

Results and discussion

Identification of satellite DNA families in the *montium* group

The TAREAN analysis in the 23 species from the *montium* group retrieved 397 clusters identified as putative satDNAs, being 245 with high confidence (HC) and 152 with low confidence (LC) (Table 1). However, after filtering the clusters using our custom satDNA filter (see Material and Methods), we selected 142 clusters for further analysis, being 124 HC and 18 LC.

We created a custom database containing consensus sequences of each one of these 142 clusters and conducted MEGABLAST-searches using each consensus sequences against our whole custom database. This analysis revealed 59 clusters that are shared among species and 83 clusters that are restricted to only one species. Therefore, the initial 142 selected clusters correspond to 101 satDNA families, which have been numbered SatDNA-1 to SatDNA-101. The general features of all 101 satDNA families can be seen in Table SM1.

Species	Subgroup*	HC satDNAs (Before filtering)	LC satDNAs (Before filtering)	HC satDNAs after satDNA filtering	LC satDNAs after satDNA filtering	Number of clusters by species after filtering
D. kanapiae	*parvula*	13	9	3	1	**4**
D. auraria	*montium*	3	10	1	1	**2**
D. triauraria	*montium*	6	5	3	0	**3**
D. asahinai	*montium*	7	8	2	1	**3**
D. rufa	*montium*	4	7	2	1	**3**
D. lacteicornis	*montium*	6	5	3	0	**3**
D. tani	*montium*	7	9	4	0	**4**
D. pectinifera	*montium*	14	5	10	0	**10**
D. punjabiensis	*punjabiensis*	14	5	5	1	**6**
D. watanabei	*punjabiensis*	7	7	2	1	**3**
D. birchii	*serrata*	15	5	7	1	**8**
D. mayri	*serrata*	15	8	13	0	**13**
D. truncata	*serrata*	11	6	5	1	**6**
D. bunnanda	*serrata*	24	6	12	2	**14**
D. serrata	*serrata*	12	9	4	2	**6**
D. bocki	*kikkawai*	10	4	6	1	**7**
D. leontia	*kikkawai*	5	7	2	2	**4**
D. jambulina	*seguyi*	8	2	5	1	**6**
D. burlai	*seguyi*	11	7	6	1	**7**
D. nikananu	*seguyi*	6	6	4	1	**5**
D. bakoue	*seguyi*	25	8	9	0	**9**
D. seguyi	*seguyi*	17	6	13	0	**13**
D. vulcana	*seguyi*	5	8	3	0	**3**
	Total:	**245**	**152**	**124**	**18**	

*According to Yassin (2018).

Table 1. SatDNA clusters identified by TAREAN after the analyses of the 23 *Drosophila* species from the *montium* group. After our custom filter, 142 clusters of satellites complied with three out of the four following parameters: c-value>0.9, p-value>0.8, high confidence and circular graph layout.

SatDNAs in the *montium* group: general structural features

Among the structural features of satellite DNAs are the monomer size and AT content. In *Drosophila*, the satDNA monomer sizes typically ranges from a few bp (≤10 bp) up to ~400 bp (Palomeque and Lorite, 2008; Melters *et al*., 2013).

The analysis of monomer sizes of the 101 selected satDNA families in species from the *montium* group revealed extensive variation, from only 4 bp (SatDNA-35 from *D. triauraria*) to 1,897 bp (SatDNA-63 from *D. burlai*). However, most satDNAs showed monomers shorter than 100 pb (Figure 3A). We found 65 satDNA families (64.4%) of <99 bp monomer length, 13 families (12.9%) of 100-199 bp, 6 families (5.9%) of 200-299 bp, 6 families (5.9%) of 300-399 bp, 5 families (4.9%) of 400-499 bp, 1 family (1.0%) of 500-599 bp, 3 families (3.0%) of 600-699 bp, 1 family (1.0%) of 800-899 bp and 1 family (1.0%) of >1000 bp. We did not identify in our data satDNAs with consensus sequences ranging from 700-799 bp and 900-999 bp. Therefore, most (89%) of the satDNAs we identified in the *montium* group are within the range of the most common monomer sizes found in *Drosophila*.

To better access the monomer size variation of the 65 satDNAs with monomer sizes shorter than 100 bp, we further subdivided this class in 10 intervals of 10 bp (Figure 3B). Most short satDNA

families have monomer sizes shorter and 10 bp (52,3%), which corresponds to 34 families. Other 15 families (23.1%) are composed of 10-19 bp monomers, 7 families (10.8%) have 20-29 bp, 3 families (4.7%) have 30-39 bp, 4 families (6.1%) have 40-49 bp, 1 family (1.5%) has 50-59 bp and 1 family (1.5%) has 90-99 bp. We did not identify satDNA consensus sequences with monomers ranging from 60 bp to 89 bp (Figure 3B).

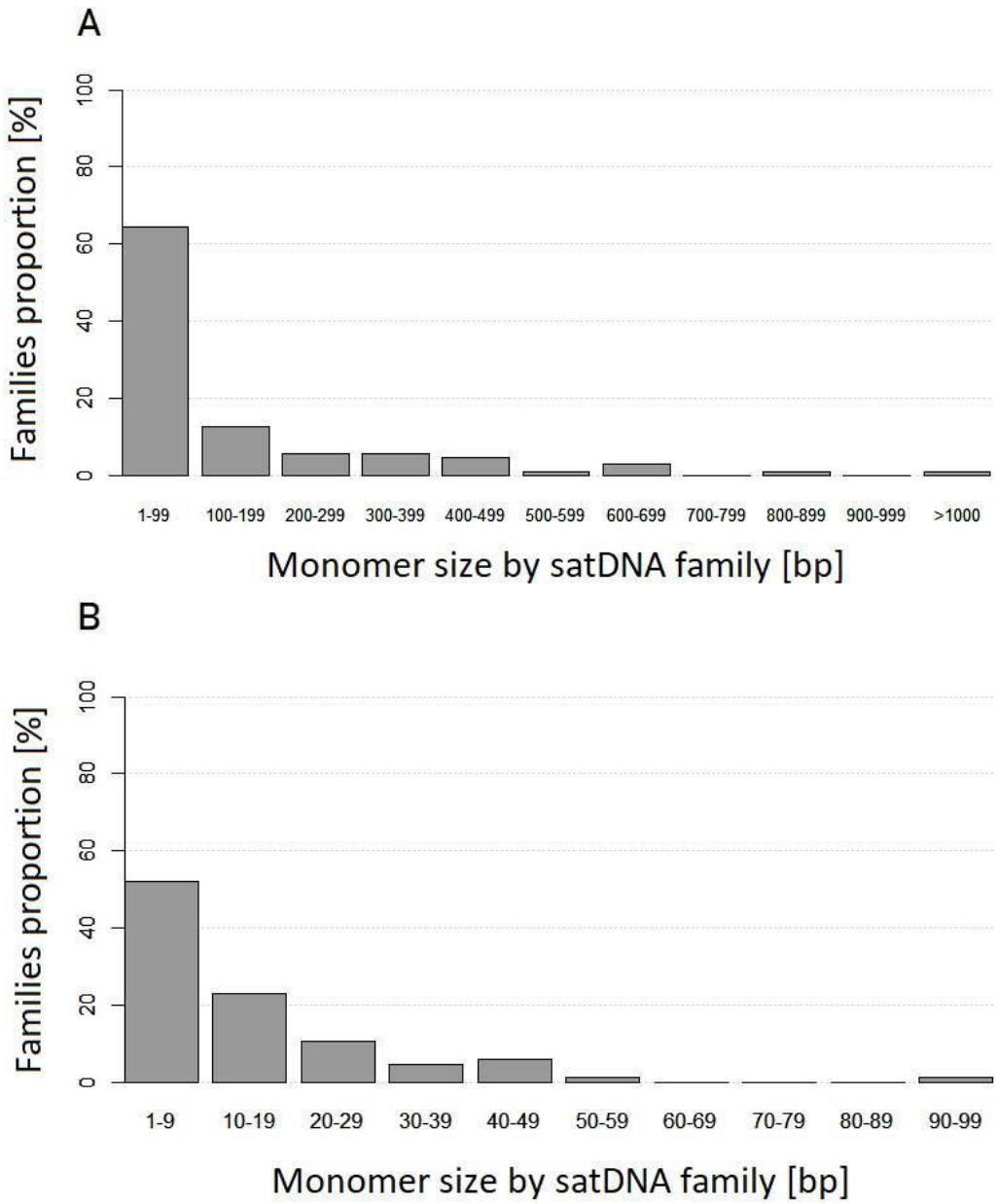


Figure 3. Monomer size by satDNA family. *A.* Monomer size of the 101 satDNA families identified. More than 60% of the total clusters retrieved are composed by sequences with 0-99bp monomer length. *B.* Monomer size of the 65 "short" satDNA families identified (0-99 bp).

Therefore, in a first view, we can infer that the 23 species genomes from the *montium* group investigated in our study are enriched with satDNAs consisting by short tandem repeats. The

presence of satDNAs with short monomer sizes in *Drosophila* is common. For example, abundant satDNA families with monomer sizes of 7 bp are found in *Drosophila virilis* (Gall *et al*., 1971, 1974) and *D. melanogaster* has several satDNAs with monomer sizes in the range of 5 bp to 10 bp (Lohe *et al*., 1993).

We used a correlation test to investigate whether there is a relationship between monomer size and copy number of the satDNA sequences. We found a strong negative correlation (ρ = -0.654; p-value = <0.01), which means that the larger the size of the monomer, the lower its number of copies in the genomes (Figure 4).

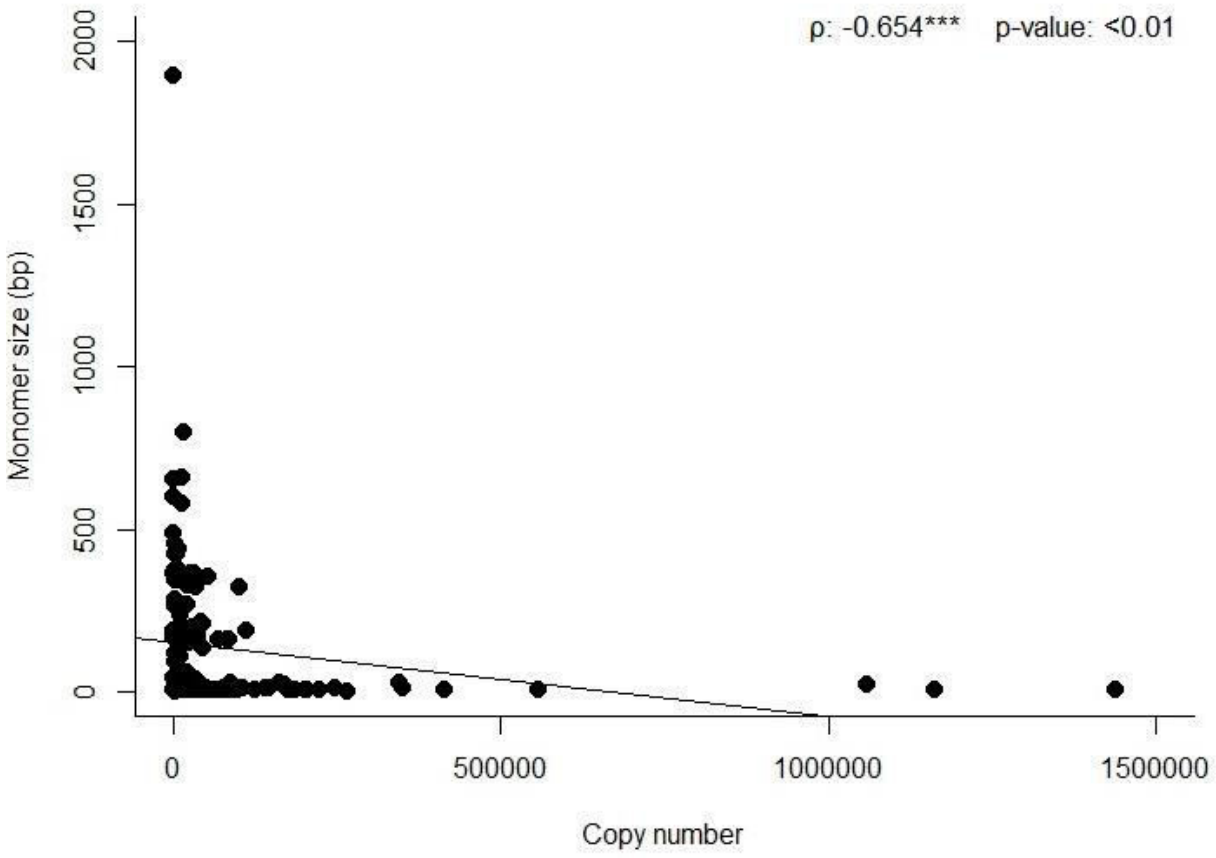


Figure 4. Correlation test between the monomer size and copy number of the 142 satDNA sequences. P-value was obtained with Spearman's correlation test. Plot graph developed in R language on RStudio Software.

In *Drosophila* and other eukaryotes, satDNAs are usually AT rich (Schmidt, 1980; Ganal and Hemleben, 1986; Palomeque and Lorite, 2008; Melters *et al*., 2013). In our collection of 101 satDNA families identified in the *montium* group, we found that 78 families have >60% of AT content on their monomer sequence (Figure 5). This number represents 76.45% of the total number of families, therefore, our findings show that satDNA sequences present in species from the *montium* group are

also mostly AT rich, as found previously for other groups and species of *Drosophila* (Gall and Atherton, 1974; Lohe *et al*., 1993).

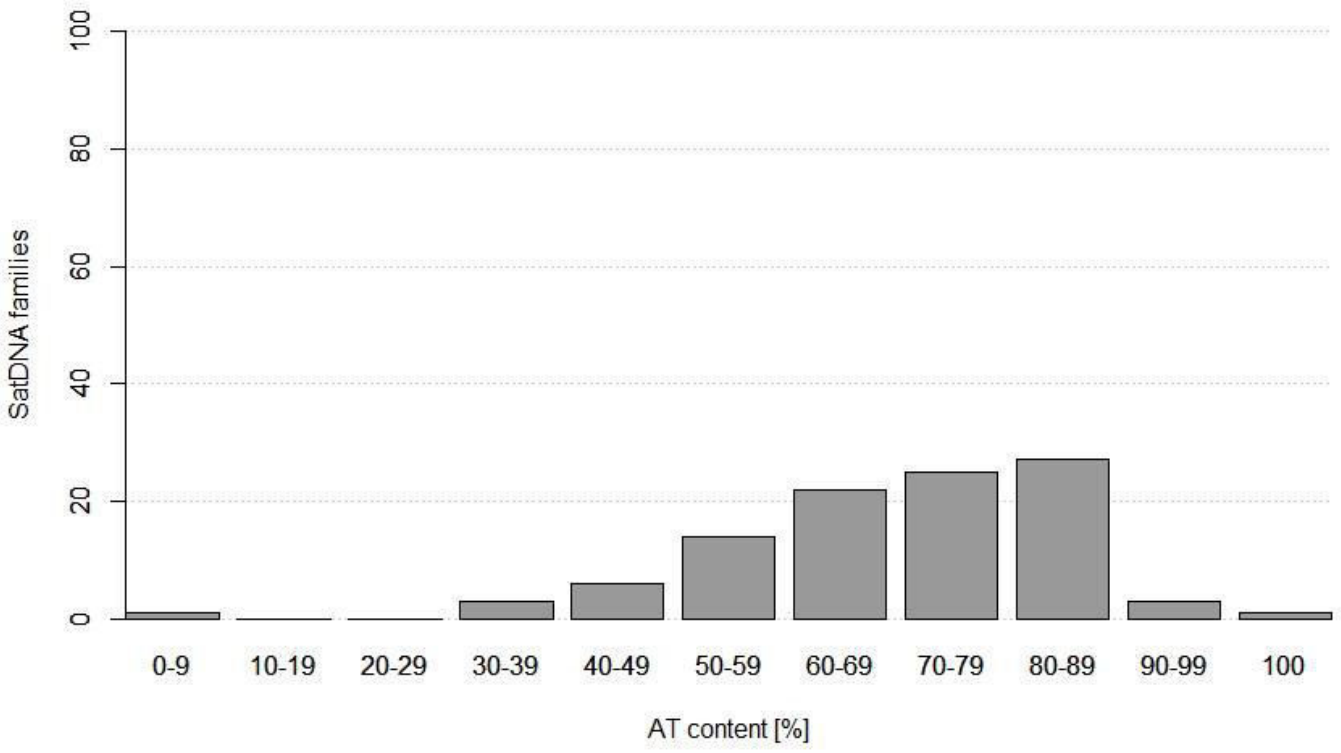


Figure 5. AT content of the 101 satDNA families identified in our study.

Other features such as the C and P values retrieved from TAREAN analyses are indispensable for a reliable satDNA identification, once both values reflect the sequence organization of tandem arrays on the genomes. Together, these values may give a good notion about the genome location of the repeats (even without cytogenetic experiments). For example: some studies showed that clusters with high genome proportion (>1%), C and P values (>0.98) and satellite probability (>0.95) may indicate satDNA sequences present on centromeric and/or pericentromeric regions of the chromosomes (Da Silva *et al*., 2020; Sena *et al*., 2020). In *Drosophila* for example, this was observed for the Sat1 satDNA family in the *virilis* group and for the pBuM satDNA in the *buzzatii* cluster (de Lima *et al*., 2017; Silva *et al*., 2019).

We analysed the C and P index for all the satDNA clusters retrieved from TAREAN and found that most clusters have >0.9 values, as expected for satellite DNAs (Table SM1). For example: the high C and P-values (>0.998) of SatDNA-6 clusters from *D. bunnanda* and *D. serrata*, SatDNA-13 in D. *vulcana* and *D. pectinifera*, SatDNA-34, SatDNA-46 and SatDNA-50 from *D. jambulina* indicate the existence of long and continuous arrays with tandem repats, what is a discriminate feature for true satDNAs.

SatDNA content and genome sizes

In *Drosophila* and many organisms, there is a positive correlation between satDNA content and genome sizes (Bosco *et al*., 2007). Furthermore, satDNAs may account for more than 20% of the genomic DNA in species from the genus, as in *D. melanogaster*, but can reach up to 70%, as in some Hawaiian *Drosophila* (Bosco *et al*., 2007; Craddock *et al*., 2016). The genome sizes in the 23 *Drosophila* studied species from the *montium* group were estimated by Bronski *et* al., (2020) and range from 155.1 Mb (*D. bocki*) to 223.4 Mb (*D. mayri*), while our estimated satDNA fraction ranges from 1.37% (*D. watanabei*) to 22.16% (*D. pectinifera*). In order to investigate if the genome sizes differences in the *montium* group are also related to the tandem repeat/satDNA fraction, we conducted correlation tests. The results are shown in the figure 6.

We performed two correlation testes with the genome sizes of the 23 genomes from the *montium* group, one with all tandem repeats identified by TAREAN (397 clusters) and another with the satDNAs (142 clusters) selected by our satDNA filter. We found positive correlations in both analyses (ρ=0.1670 and ρ=0.222) but with weak statistical support (p-value>0.05) (Figure 6). Therefore, these results suggest that satDNAs are not the main genomic components responsible for genome size variation among species from the *montium* group.

Bronski *et al*., (2020) found a strong positive correlation between the estimated genome sizes and the whole repetitive content across all the 23 *montium* genomes. Specially, as an example, it is important to highlight that we also found the two biggest tandem repeat/satDNA fractions in the *D. pectinifera* and *D. mayri* data (the two species with the largest genomes), as already showed by Bronski et al., (2021) (Figure 6A and 6B). Additionally, *D. kanapiae*, D*. bocki* and *D. leontia* have short genomes and low satDNA fraction as well (Figure 6B). In this context, our findings are in accordance with the previous results found by Bronski *et al*., (2021), that is, in most cases the bigger the genome, the bigger the tandem repeat/satDNA fraction.

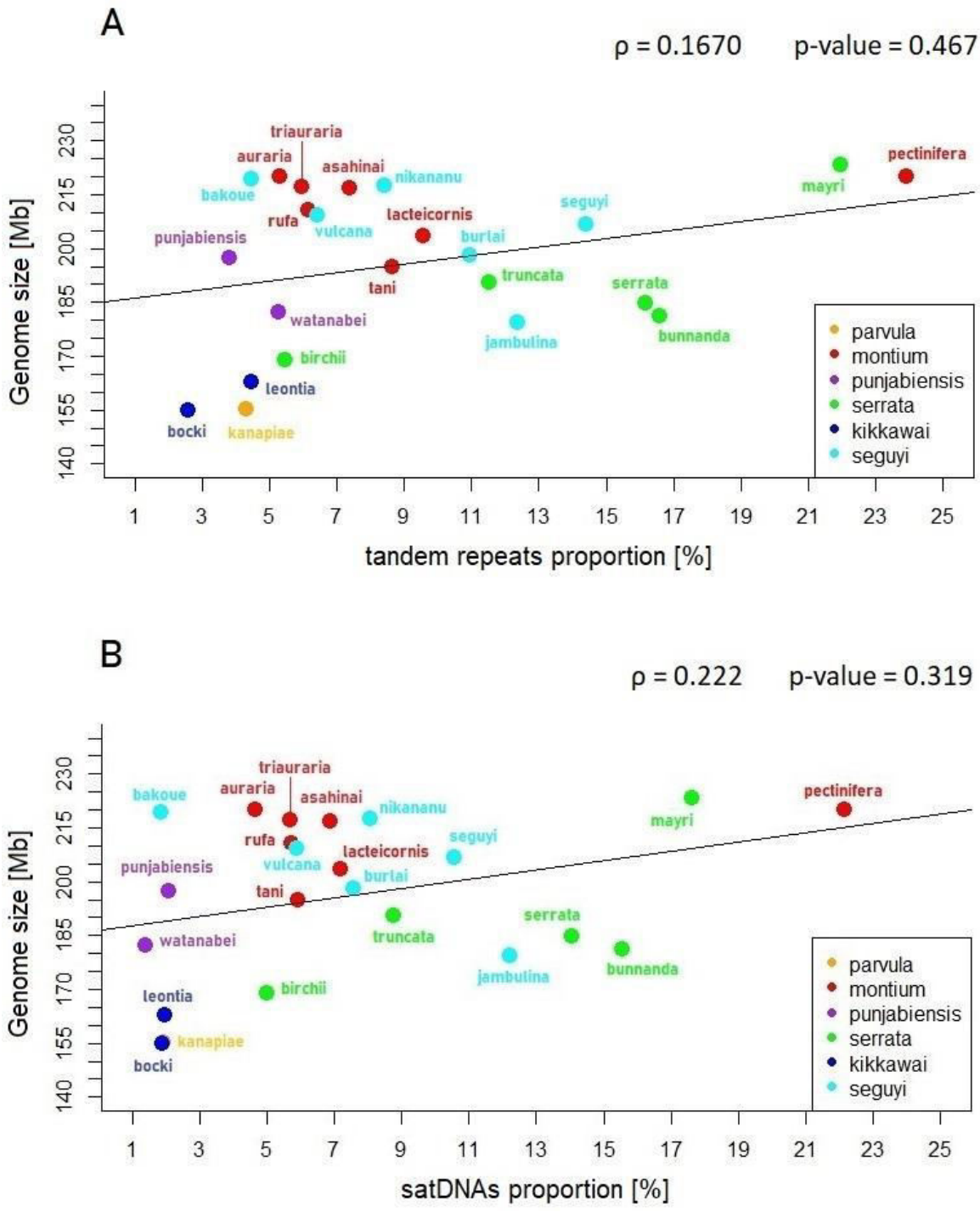


Figure 6. Correlation tests. *A.* Correlation between genome size and genome proportion of all tandem repeats (397 clusters). *B.* Correlation between genome size and satDNAs genome proportion (142 clusters). P-values were obtained with Spearman's correlation test. The colors are based on subgroup classification proposed by Yassin (2018).

SatDNAs and phylogenetic relationships within species from the *montium* group

From all 101 satDNA families identified by TAREAN in species from the *montium* group, only 17 were found present in two or more species (Figure 7). No satDNA family was found present in all species from the group. This result illustrates the fast evolutionary dynamics of satDNAs. In this context, it is worth mentioning that 84 satDNA families (83% of the total) have been found restricted to single species.

Below, we describe how each one of the 17 satDNA families shared by at least two species contribute for the recently proposed classification of subgroups within the *montium* group and also for the phylogenetical relationships between some species. We did not use SatDNA-9, SatDNA-13, SatDNA-15, SatDNA-16 and SatDNA-17 for phylogenetic inferences due to the short sizes of their monomers (≤10bp) and the high (100%) interspecific nucleotide identity between the copies.

For these 17 satDNA families, we analyzed the graph layouts retrieved by TAREAN. According to Novák *et al*., (2017), the shapes of the graphs reflect the genomic organization and sequence variability of the clusters. The shapes range from linear structures (found for dispersed elements, as TEs) to highly circular or globular structures, common for tandemly repeated sequences, such as satDNAs. Therefore, satDNA families which are composed of more variable sequences organized on interspersed arrays may present circular but not globular satDNA-like graphs. In our TAREAN analyses, we found circular satDNA-like graphs in almost all families, except for the SatDNA-12 family. For example, circular satDNA-like graph layouts of the SatDNA-1 family is shown in the Figure SM3.

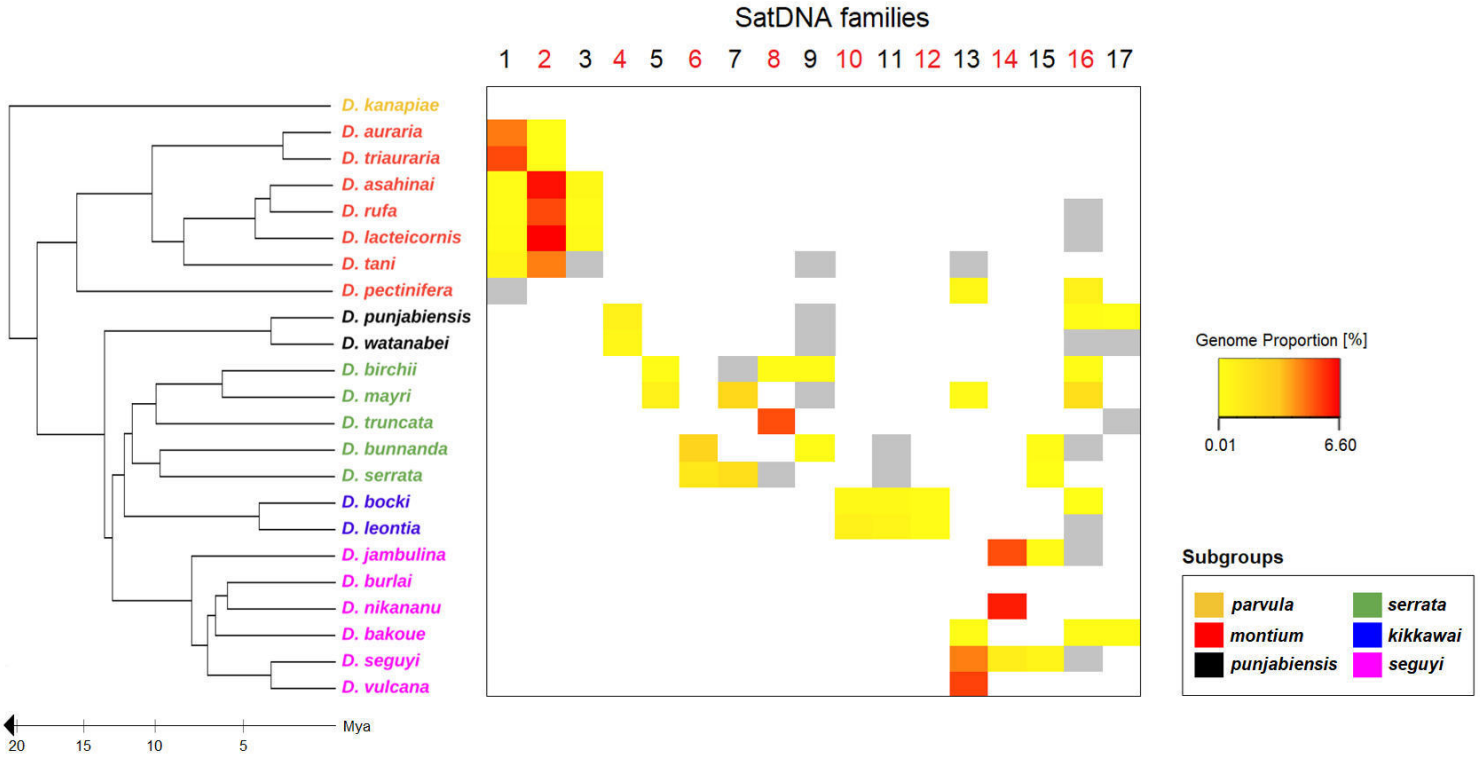


Figure 7. Heatmap showing the genome proportion for each cluster of the satDNA families identified

by TAREAN after filtering. The phylogenetic tree was reconstructed according to Yassin (2018). The exact genome proportion values for each satDNA are described in Table SM1. Gray squares indicate the presence of homologous sequences identified by MEGABLAST searches on NCBI database.

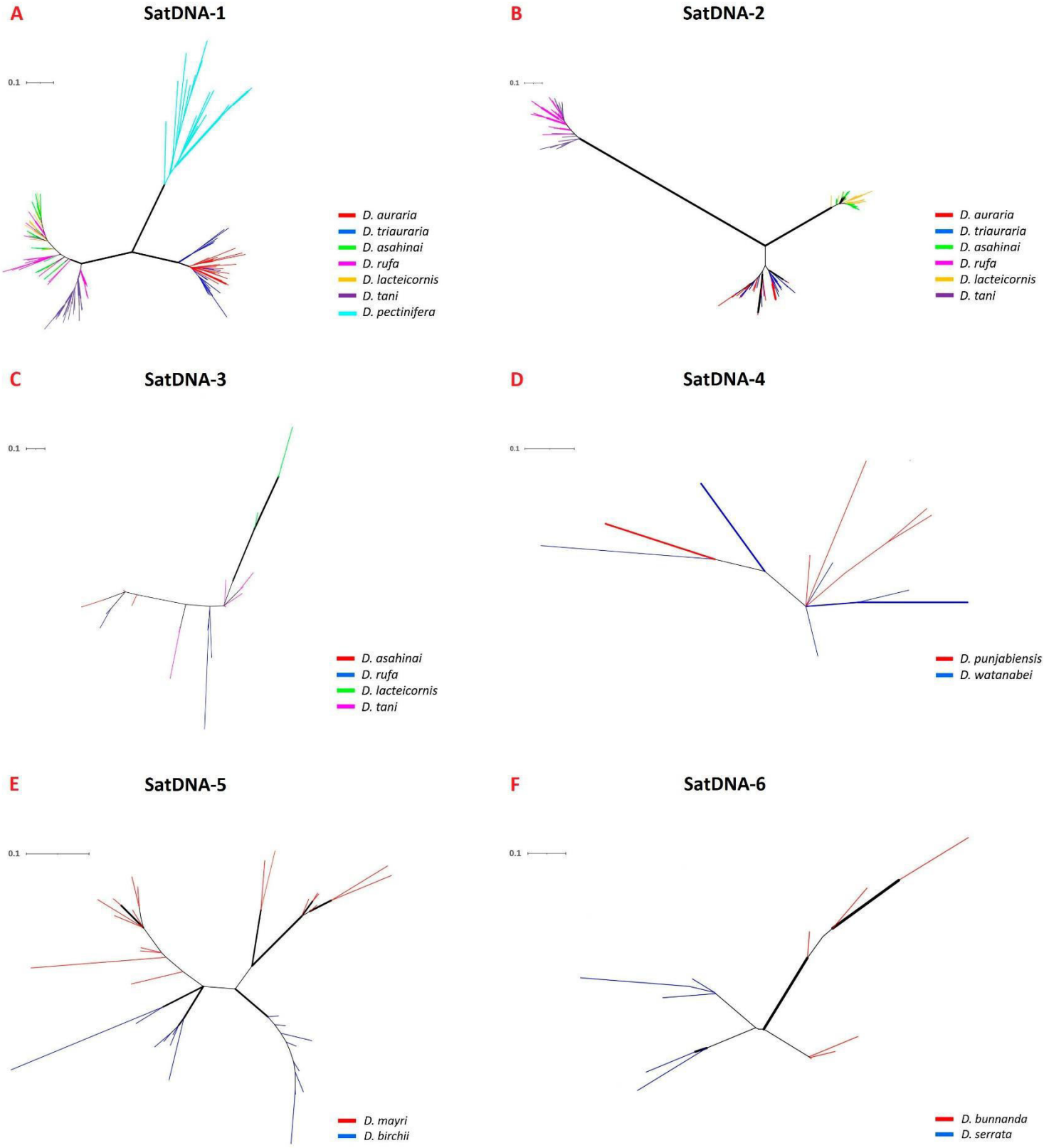
SatDNA-1

The SatDNA-1 family has been identified by TAREAN in all species belonging to the *montium* subgroup, except *D. pectinifera,* a species that occupies a basal position within the subgroup (Figure 7). However, we identified SatDNA-1 homologous sequences in *D. pectinifera* using MEGABLAST searches against its genome. We have not found SatDNA-1 outside the *montium* subgroup.

The SatDNA-1 satDNA family has monomers of approximately 370 bp showing an AT content of 74.27% on average. In the TAREAN results, SatDNA-1 displayed circular graph layouts, a feature that suggest its classification as being a true satellite DNA (Figure SM3).

In the *D. auraria*/*D. triauraria* branch, SatDNA-1 genomic proportions are 4.6% and 5.4%, respectively. In the other species branch (*D. asahinai*, *D. rufa*, *D. lacteicornis* and *D. tani*), the SatDNA-1 genomic proportions are much smaller, 0.68%, 0.37%, 0.77% and 0.72%, respectively (Table SM1). Therefore, satDNA genomic proportion supports the existence of these two branches within the *montium* subgroup.

The ML tree shows SatDNA-1 sequences grouped into three main branches. One branch is made of SatDNA-1 copies from *D. pectinifera*, another branch is made by copies of *D. auraria* and *D. triauraria,* while the third branch is composed by copies from *D. tani*, *D. rufa*, D. *lacteicornis* and *D. asahinai* (Figure 8A). These phylogenetic relationships are in accordance with recent studies by Yassin (2018) and Conner *et al*., (2021) (Figure 1).


A
SatDNA-1
0.1
D. auraria
D. triauraria
D. asahinai
D. rufa
D. lacteicornis
D. tani
D. pectinifera
B
SatDNA-2
0.1
D. auraria
D. triauraria
D. asahinai
D. rufa
D. lacteicornis
D. tani
C
SatDNA-3
0.1
D. asahinai
D. rufa
D. lacteicornis
D. tani
D
SatDNA-4
0.1
D. punjabiensis
D. watanabei
E
SatDNA-5
0.1
D. mayri
D. birchii
F
SatDNA-6
0.1
D. bunnanda
D. serrata

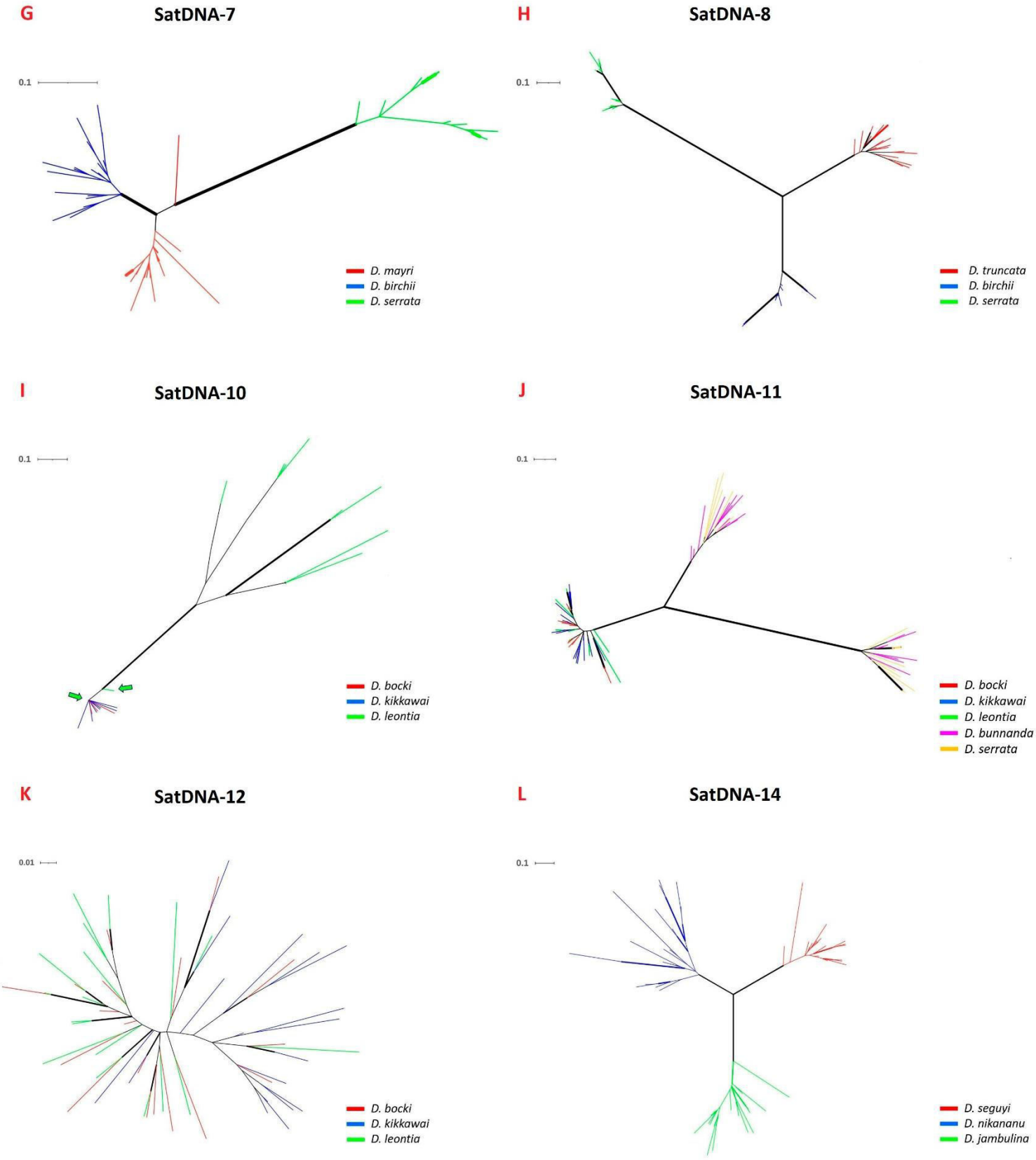


Figure 8. Maximum Likelihood trees based on satDNAs families. Ticker branches illustrate bootstrap values >0.7. The best-fit evolutionary model was selected in the multiple nucleotide alignments of each

satDNA. *A.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-1 sequences. *B.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-2 sequences. *C.* Maximum-Likelihood (T92) phylogeny of SatDNA-3 sequences. *D.* Maximum-Likelihood (JC) phylogeny of SatDNA-4 sequences. *E.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-5 sequences. *F.* Maximum-Likelihood (JC) phylogeny of SatDNA-6 sequences. *G.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-7 sequences. *H.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-8 sequences. *I.* Maximum-Likelihood (HKY) phylogeny of SatDNA-10 sequences. Green arrows indicate the presence of *D. leontia* copies (4 in total). *J.* Maximum-Likelihood (T92+G) phylogeny of SatDNA-11 sequences. *K.* Maximum-Likelihood (K2+G) phylogeny of SatDNA-12 sequences. *L.* Maximum-Likelihood (HKY+G) phylogeny of SatDNA-14 sequences. Scale bar represents 0.1 (10% differences between sequences) for all satDNA families but SatDNA-12 (0.01 = 1%).

SatDNA-2

Similarly to SatDNA-1, the SatDNA-2 family has been identified in all species belonging to the *montium* subgroup, except *D. pectinifera* (Figure 7). MEGABLAST searches did not retrieve homologous sequences in *D. pectinifera* or in any other *Drosophila* species.

The SatDNA-2 family has monomers of approximately 165 bp long and AT content of 65.03% on average. As found for the SatDNA-1 family, SatDNA-2 also presented circular satDNA graph layouts in the TAREAN results.

In the *D. auraria*/*D. triauraria* branch, SatDNA-2 genomic proportions are 0.021% and 0.015%, respectively. In the *D. asahinai*/*D. rufa*/*D. lacteicornis/D. tani* branch, the genomic proportion is much higher, 6.30%, 5.40%, 6.60% and 4.50%, respectively (Table SM1). Interestingly, this pattern of high and low genomic proportion is the opposite to what we found for SatDNA-1, indicating that a substantial turnover of SatDNA-1 and SatDNA-2 amount happened in the genome of species belonging to these two branches: while SatDNA-1 is more abundant in *D. auraria* and *D. triauraria* and less abundant in *D. asahinai*, *D. rufa*, *D. lacteicornis* and *D. tani,* SatDNA-2 is more abundant in *D. asahinai*, *D. rufa*, *D. lacteicornis* and *D. tani* and less abundant in *D. auraria* and *D. triauraria.*

The ML tree shows *D. auraria* and *D. triauraria* copies clustered in a single branch (as shown for SatDNA-1), copies from *D. asahinai* and *D. lacteicornis* forming a second branch and copies from *D. tani* e *D. rufa* forming a third branch (Figure 8B). This result differs from SatDNA-1 on clustering of *D. asahinai*/*D. rufa*/*D. lacteicornis/D. tani* branch. The ML tree topology for SatDNA-2 is in accordance with phylogenetic trees made by Conner *et al*., (2021) (Figure 1B), where gene sequences from *D. asahinai* are more similar to *D. lacteicornis* sequences than to *D. rufa* sequences. For this satDNA family, *D. rufa* copies are more similar to *D. tani* copies.

SatDNA-3

SatDNA-3 has been identified in *D. asahinai*, *D. rufa*, *D. lacteicornis* by TAREAN and in *D. tani* by MEGABLAST searches (Figure 7).

This satDNA family is composed of short tandem repeat sequences (17 bp) and has 76.48%

of average AT content. The SatDNA-3 clusters presented circular satDNA-layouts, as found for SatDNA-1 and SatDNA-2 in the TAREAN results.

Unlike SatDNA-1 and SatDNA-2, this satDNA is less abundant. The SatDNA-3 genomic proportion is 0.36% in *D. asahinai*, 0.31% in *D. lacteicornis* and 0.15% in *D. rufa* (Table SM1). These values are similar, what is expected for closely related species.

The ML tree with SatDNA-3 sequences (Figure 8C) showed no species-specific branches, indicating very low divergence across species. Therefore, the data from this satDNA family is in accordance with the close phylogenetic relationships of *D. asahinai*, *D. rufa*, *D. lacteicornis* and *D. tani* found by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

SatDNA-4

According to the TAREAN analyses and MEGABLAST searches, SatDNA-4 is restricted to *D. punjabiensis* and *D. watanabei*, two species from the *punjabiensis* subgroup (Figure 7).

This satDNA family is composed of monomers 20 bp long and its average AT content is 90.0%. For both species, the SatDNA-4 clusters retrieved by TAREAN displayed circular satDNA-like graph layouts.

The SatDNA-4 genome proportion varies ~2x between species, with *D. punjabiensis* presenting 0.92% and *D. watanabei* presenting 0.49% (Table SM1).

The ML tree with SatDNA-4 sequences showed no evident species-specific branches, indicating low inter-specific divergence (Figure 8D). The presence of this restricted satDNA in *D. punjabiensis* and *D. watanabei* is in accordance to their classification into the *punjabiensis* subgroup (Figure 1). The correct position of the *punjabiensis* subgroup inside the *montium* group phylogeny is controversial according to Yassin (2018) and Conner *et al*., (2021) (Figure 1). In this case, SatDNA-4 distribution in the *montium* group does not help to elucidate the correct position of the *punjabiensis* clade (Figure 7).

SatDNA-5

The SatDNA-5 family was detected by TAREAN in *D. birchii* and *D. mayri,* two species from the *serrata* subgroup (Yassin, 2018). We did not find this satDNA in the others species via MEGABLAST searches (Figure 7).

The sizes of SatDNA-5 consensus sequences from our TAREAN analyses are more divergent in this family (190 bp in *D. mayri* and 121 bp in *D. birchii*) and the average AT content of monomer is 75.15%. As found for the previous families, SatDNA-5 also presented circular satDNA-like graph layouts in the TAREAN results.

The SatDNA-5 genome proportions retrieved are very different in the two species: 0.92% in *D. mayri* and 0.14% in *D. birchii* (Table SM1).

The ML tree shows four branches with two subdivisions from both species (Figure 8E). This topology may indicate presence of satDNA subfamilies in *D. mayri* and *D. birchii*. The fact that

SatDNA-5 is restricted to *D. mayri* and *D. birchii* supports the close phylogenetical relationship between these species within the *serrata* subgroup, as previously suggested by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

SatDNA-6

SatDNA-6 is a satDNA family identified by TAREAN which is only present in *D. bunnanda* and *D. serrata*, two species from the *serrata* subgroup (Figure 1). We did not identify this satDNA in the other species with MEGABLAST searches (Figure 7).

This satDNA is composed of monomers ~30 bp long and the average AT content is 55.61%. The SatDNA-6 also displayed circular satDNA-like graph layouts in the TAREAN results.

SatDNA-6 is an abundant satDNA family, with genomic contribution of 2.7% in *D. bunnanda* and 1.4% in *D. serrata* (Table SM1).

The ML tree with SatDNA-6 sequences is shown in the figure 8F and, as for SatDNA-5, the topology indicates the presence of two subfamilies in each species. The exclusive presence of SatDNA-6 in *D. serrata* and *D. bunnanda* is in accordance to the close phylogenetical relationship between these species within the *serrata* subgroup, as previously suggested by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

SatDNA-7

SatDNA-7 is a satDNA family identified by TAREAN in *D. mayri* and *D. serrata* and with MEGABLAST searches in *D. birchii* (Figure 7). We found this satDNA restricted to the genomes of species from the *serrata* subgroup, with the exception of *D. bunnanda* and *D. truncata*.

The SatDNA-7 consensus sequence has 150 bp in *D. mayri* and 153 bp in *D. serrata* and the average AT content is 72.58%. Additionally, SatDNA-7 presented circular satDNA-like graph layouts in TAREAN results.

This is an abundant satDNA family which displays 2.0% of genome proportion in *D. serrata* and 2.5% in *D. mayri* (Table SM1). According to Yassin (2018), *D. serrata* and *D. mayri* are species phylogenetically separated about 11 Mya, and even with this long divergence time, they have close genomic proportions of SatDNA-7. Interestingly, this result is the opposite found for SatDNA-6, a satDNA family present in *D. bunnanda* and *D. serrata* (separated about 9 Mya), whose genome proportions are more divergent between the species (Figure 7).

The ML tree shows SatDNA-7 sequences mainly clustered in species-specific branches, with sequences from *D. mayri* and *D. birchii* more closely connected (Figure 8G). Also, *D. serrata* copies are clustered in a different and distant branch reinforcing the basal position of this species. Our findings with SatDNA-7 are in accordance with the phylogenetic relationships between these species proposed by Yassin (2018) and Conner *et al*., (2021).

SatDNA-8

SatDNA-8 is a satDNA family identified by TAREAN in *D. birchii* and *D. truncata* and, with MEGABLAST searches, also in *D. serrata* (Figure 7). Therefore, as found for SatDNA-5, SatDNA-6 and SatDNA-7, this is another satDNA restricted to species from the *serrata* subgroup.

TAREAN identified two SatDNA-8 monomer variants in *D. truncata*, one with 198 bp and another 194 bp, while in *D. birchii*, only a 182 bp long monomer were found (Table SM1). The average AT content of SatDNA-8 is 72.06% and all the clusters retrieved by TAREAN showed a circular satDNA-like graph layout.

The SatDNA-8 family is more abundant in *D. truncata* (5.3%) than in *D. birchii* (0.21%), what suggests: 1) a loss of copies in *D. birchii* after the separation from the *D. truncata* clade (~ 9 Mya) or 2) a recent gain of copies in the *D. truncata* clade. The absence of this satDNA in the others species from the subgroup (*D. mayri* and *D. bunnanda*) can be explained by independent events of loss of this satDNA family in these lineages.

The ML tree containing SatDNA-8 sequences shows three distinct species-specific branches (Figure 8H). The branches leading to SatDNA-8 from *D. birchii* and *D. truncata* are more closely connected in relation to the *D. serrata* branch. The tree topology of SatDNA-8 is in agreement with the phylogenetic relationships between these three species proposed by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

SatDNA-9

The SatDNA-9 family has been identified in *D. birchii* and *D. bunnanda* by TAREAN, and in *D. mayri*, *D. watanabei*, *D. punjabiensis* and *D. tani* by MEGABLAST searches (Figure 7). Thus, this satDNA family was found in species from the *montium*, *punjabiensis* and *serrata* subgroup.

SatDNA-9 is a satDNA family with monomers of 6 bp and average AT content of 83.34%. As found for most families, SatDNA-9 clusters retrieved by TAREAN also showed a circular satDNA-like graph layout.

The genomic contribution of SatDNA-9 in *D. birchii* and *D. bunnanda* is 0.130% (Table SM1). This satDNA is conserved in the *punjabiensis* subgroup (*D. punjabiensis* and *D. watanabei*) and in the *D. birchii*/*D. mayri* clade (Figure 1). Therefore, the distribution of SatDNA-9 is mostly in accordance with the previous phylogenies proposed by Yassin (2018) and Conner *et al*., (2021).

SatDNA-10

The SatDNA-10 family has been identified in *D. bocki* and *D. leontia* by TAREAN (Figure 7). Additionally, we found this satDNA family in *D. kikkawai* with MEGABLAST searches. Thus, this TR family is only present in the *kikkawai* subgroup.

The TAREAN analyses revealed that the consensus of SatDNA-10 is 41 bp long, with an average AT content of 65.86%. Moreover, this satDNA family shows circular satDNA-like graph layouts.

The SatDNA-10 genome proportions retrieved by TAREAN are 0.41% in *D. bocki* and 0.89%

in *D. leontia* (Table SM1). Although this satDNA is not too abundant (<1.0%), the genome proportion in *D. leontia* is 2x higher than in *D. bocki*.

The ML tree with SatDNA-10 sequences is shown in the figure 8I. Copies from *D. bocki* and *D. kikkawai* are clustered together with few copies from *D. leontia*. However, most copies of *D. leontia* are clustered in another branch, which reveals a lower similarity between satDNA copies from this species when compared to the *D. kikkawai* and *D. bocki* copies. The phylogeny proposed by Yassin (2018) clustered these species in a polytomous clade, whereas Conner *et al*., (2021) proposed *D. bocki* and *D. leontia* as sister species and *D. kikkawai* on a basal position in the clade (Figure 1). The SatDNA-10 shows a different history, where repeats from *D. kikkawai* and *D. bocki* are most closely related when compared to *D. leontia* (Figure 8I).

SatDNA-11

The SatDNA-11 family has been identified in species from the *kikkawai* (*D. bocki*, *D. leontia*, *D. kikkawai*) and *serrata* (*D. bunnanda* and *D. serrata*) subgroups (Figure 7). Initially, we identified the clusters via TAREAN analyses in the genomes of *D. bocki* and *D. leontia*. The presence of this satDNA family in other species was confirmed by MEGABLAST searches.

The SatDNA-11 consensus sequence is 109 bp long and the average AT content is 73.40%. Additionally, SatDNA-11 shows circular satDNA-like graph layouts.

The SatDNA-11 family contributes with 0.38% of the genome in *D. bocki* and 0.66% in *D. leontia* (Table SM1). The presence of this satDNA family in species from the *serrata* subgroup may be an indicative of the proximal phylogenetic relation between the *kikkawai* and *serrata* subgroups, as previously proposed by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

The ML tree with SatDNA-11 sequences shows three well-defined and distinct branches (Figure 8J). Copies from *D. bocki*, *D. leontia* and *D. kikkawai* are grouped together whereas copies from *D. serrata* and *D. bunnanda* are grouped in two different branches. The presence of two different branches for *D. bunnanda* and *D. serrata* copies may suggest the presence of SatDNA-11 subfamilies.

SatDNA-12

Similarly to SatDNA-10, SatDNA-12 has been identified in species from the *kikkawai* subgroup (Figure 7). The identification of SatDNA-12 was based in our TAREAN analyses for *D. bocki* and *D. leontia* and with MEGABLAST searches for *D. kikkawai.*

This satDNA family has monomers of 489 bp and average AT content of 30.48%. Regarding the graph layouts, the TAREAN analyses retrieved linear structures, indicating the presence of dispersed repeats.

The SatDNA-12 genomic proportion is 0.16% for *D. bocki* and 0.10% for *D. leontia* (Table SM1).

The ML tree with SatDNA-12 sequences shows branches made by repeats from different

species, indicating no inter-specific divergence (Figure 8K). Therefore, based on SatDNA-12, it is not possible to infer the phylogenetic relationships between these three species. On the other hand, the presence of this satDNA family restricted to *D. bocki*, *D. leontia* and *D. kikkawai* confirm the close phylogenetic relationship between these three species from the *kikkawai* subgroup, as proposed by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

SatDNA-13

The SatDNA-13 family has been identified in *D. pectinifera*, *D. mayri*, *D. bakoue*, *D. seguyi* and *D. vulcana* by TAREAN and in *D. tani* by MEGABLAST searches. Therefore, this satDNA is found in three subgroups but in not all species from these subgroups (Figure 7). According to our results, this TR does not show a conserved distribution in the *montium* group following the phylogenies proposed by Yassin (2018) and Conner *et al*., (2021) (Figure 1).

The SatDNA-13 monomer is 8 bp long and the average AT content is 85.0%. Additionally, SatDNA-13 graphs retrieved by TAREAN show circular satDNA-like shapes.

The SatDNA-13 is very abundant in *D. seguyi* and *D. vulcana,* comprising ~ 5% of the total genomic DNA (Table SM1). However, we also found SatDNA-13 in a smaller genomic content (<1.0%) in *D. mayri*, *D. bakoue* and *D. pectinifera*, which indicates that this sequence expanded in the *D. seguyi*/*D. vulcana* clade.

SatDNA-14

The SatDNA-14 family has been identified by TAREAN in *D. jambulina*, *D. nikananu* and *D. seguyi*, species from the *seguyi* subgroup (Figure 7). We did not find this satDNA in species from the other subgroups via MEGABLAST searches.

For SatDNA-14, TAREAN retrieved three *D. nikananu* monomer variants (probably representing three subfamilies), with 188 bp, 191 bp and 207 bp long, one variant of 208 bp long in *D. jambulina* and one variant of 178 bp long in *D. seguyi*. The average AT content of these SatDNA-14 monomers is 69.15% and the clusters retrieved circular satDNA-like graph layouts.

The genomic proportions of the three *D. nikananu* SatDNA-14 variants are 2.4% (188 bp), 2.7% (191 bp) and 0.99% (207 bp) (Table SM1). Together, these variants make at least 6.09% of the *D. nikananu* genome. In *D. jambulina*, the SatDNA-14 displays 5.3% of genomic proportion and in *D. seguyi* the family displays 1.1% of genomic proportion.

The ML tree with SatDNA-14 sequences shows three species-specific long branches (Figure 8L). The *seguyi* subgroup topology in the phylogenetic trees proposed by Yassin (2018) and Conner *et al*., (2021) are quite different (Figure 1). However, according to our SatDNA-14 ML tree results, it is not possible to infer the kinships between the species based on these previous studies.

SatDNA-15

The SatDNA-15 family was identified by TAREAN in *D. bunnanda*, *D. serrata*, *D. jambulina* and *D. seguyi* (Figure 7). We did not find this satDNA in the other species by MEGABLAST searches.

Accordingly, this family was only found is species from the *serrata* and *seguyi* subgroups but not in all species from each subgroup. Thus, the distribution of this satDNA does not reflect the phylogenetic relationships proposed by Yassin (2018) and Conner *et al*., (2021).

The SatDNA-15 family is composed of short 9 bp long monomers, whose average AT content is 88.25%. Additionally, the clusters retrieved by TAREAN show circular satDNA-like graph layouts.

For this satDNA, the genomic proportions range from 0.040% in *D. serrata* to 0.280% in *D. seguyi*, showing a high variation (up to 7x) between the species (Table SM1).

SatDNA-16

The SatDNA-16 family has been identified in 13 of the 23 analysed species: *D. pectinifera*, *D. punjabiensis*, *D. birchii*, *D. mayri*, *D. bocki* and *D. bakoue* by TAREAN and in *D. rufa*, *D. lacteicornis*, *D. watanabei*, *D. bunnanda*, *D. leontia*, *D. jambulina* and *D. seguyi* by MEGABLAST searches (Figure 7).

In almost all species, the monomers are 8 bp long, with the exception of *D. pectinifera* (10 bp). The average AT content is 87.91% and the clusters present circular satDNA-like graph layouts.

The genomic proportion of SatDNA-16 is 2% in *D. mayri* and 0.930% in *D. pectinifera*. In the other species, SatDNA-16 is less abundant, ranging from 0.026% in *D. bocki* to 0.170% in *D. bakoue* (Table SM1). This is an interesting satDNA which is expanded in some species and restricted in others. Furthermore, SatDNA-16 family is the main example of ubiquitous satDNA found in different species and subgroups. Thus, the distribution of this satDNA family does not follow the species relationships within and between subgroups previously proposed by Yassin (2018) and Conner *et al*., (2021).

SatDNA-17

The SatDNA-17 family has been identified by TAREAN in *D. bakoue* and *D. punjabiensis* and with MEGABLAST searches in *D. watanabei* and *D. truncata* (Figure 7). Therefore, SatDNA-17 is restricted to few species from the *punjabiensis*, *serrata* and *seguyi* subgroups.

The SatDNA-17 family is composed of monomers 9 bp long, with average AT content of 88.89%. Additionally, this satDNA display circular satDNA-like graph layouts.

The genomic proportion of SatDNA-17 is 0.210% in *D. bakoue* and 0.061% in *D. punjabiensis* (Table SM1). Compared to SatDNA-16, which is also a short satDNA family, SatDNA-17 is present in few species (only four), from different and distant subgroups (*seguyi*, *serrata* and *punjabiensis*). Thus, as found for SatDNA-16, the distribution of this satDNA family does not follow the phylogenetic relationships within and between subgroups previously proposed by Yassin (2018) and Conner *et al*., (2021).

Abundant satDNAs sequences in the *montium group* share homology with Helitron

transposable elements

Helitrons are eukaryotic transposable elements (TEs) present in many organisms, from protists to animals and plants (Feschotte and Pritham, 2007). These copy-and-paste repetitive sequences move throughout the genomes by a rolling-circle-like replication mechanism, via a single-stranded DNA intermediate (Feschotte and Wessler, 2001; Kapitonov and Jurka, 2001).

Over the last 20 years, new variants of Helitrons have been discovered, so that they have been currently classified into 4 main groups: Helentrons, Proto-Helentron, Helitron 2 and DINEs. The DINEs (*Drosophila* INterspersed Elements) are abundant repetitive elements found in some drosophilids, as in *D. melanogaster*, for example (Thomas and Pritham, 2015). Generally, DINE sequences are composed by two conserved blocks (A and B) separated by central tandem repeats (CTRs), although some variants do not have this organization.

Recently, Dias *et al*., (2015) identified a DINE variant they named DINE-TR1 in 13 *Drosophila* species and *Bactrocera tryoni* (Acalyptratae, Diptera). The DINE-TR1 has in its structure the conserved blocks A and B and CTRs of ~150 bp. Interestingly, the CTRs were found amplified leading to long satDNA-like arrays in three *Drosophila* species: *D. virilis*, *D. americana* and *D. biarmipes*. According to Dias *et al*., (2015), DINE-TR1 and its CTRs may represent a potential source for satDNA emergence, as these expansion of CTRs occurred independently at least twice in the *Drosophila* genus.

We also investigated if the 101 TR satDNA families identified in the present work share homologous regions with transposable elements, specially Helitrons. For this purpose, we looked for homologous TE sequences deposited in the CENSOR database (Kohany *et al*., 2006) (on Repbase) and in a custom database made by 41 consensus sequences of *Drosophila* Helitrons. The results are shown in Table SM1. We found that 12 satDNA families share highly similar sequences to *Drosophila* Helitrons, being them: SatDNA-1, SatDNA-7, SatDNA-8, SatDNA-14, SatDNA-20, SatDNA-22, SatDNA-41, SatDNA-67, SatDNA-79, SatDNA-81, SatDNA-84 and SatDNA-91 (Table SM2). Below, we describe in more details the results found for the SatDNA-7, which is another interesting case of CTR expansion towards satDNA-like arrays.

SatDNA-7 is an abundant satDNA present in *D. serrata*, *D. mayri* and *D. birchii* (Figure 7). We found SatDNA-7 consensus sequences sharing homology with different DINEs sequences, such as: Helitron-N1, Helitron-N2, Helitron-2, Helitron2N and Helitron2N1. These DINEs differ from each other mainly by the CTR size. One of these hits is the Helitron-N2 of *D. biarmipes,* which has been classified as a DINE-TR1 with CTRs made of ~162 bp copies between blocks A and B (Dias *et al*. 2015; Figure 9A). This DINE-TR1 from *D. biarmipes* is the same that has experience amplification of its CTRs.

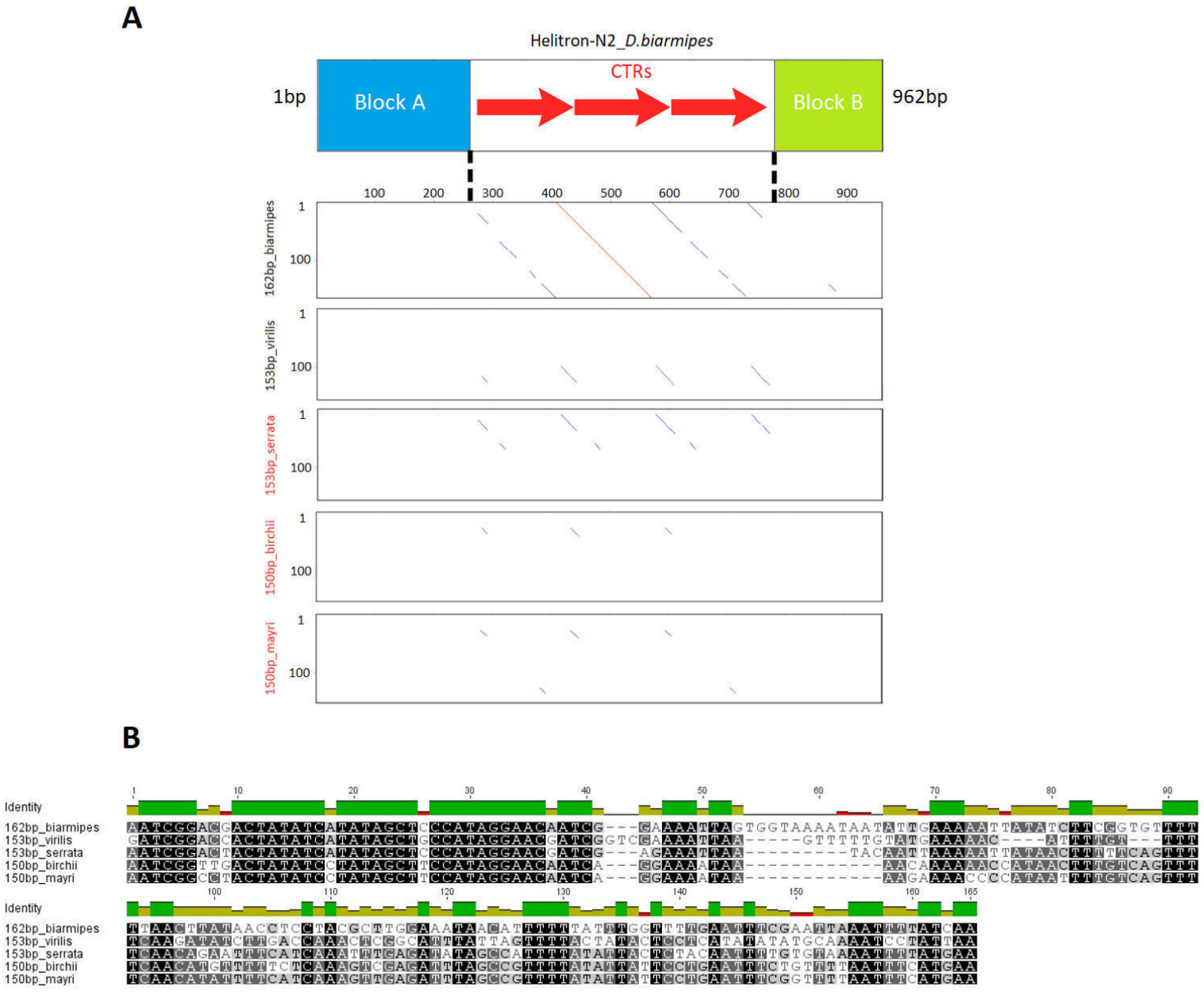


Figure 9. The SatDNA-7 family is composed of DINE-TR1 CTR sequences in *D. serrata*, *D. birchii* and *D. mayri*. *A.* Dotplot illustrating the nucleotide sequence conservation between *D. biarmipes* CTRs with *D. virilis*, *D. serrata*, *D. birchii* and *D. mayri* CTR sequences. *B.* Nucleotide sequence alignment of the three CTR consensus found in *D. serrata*, *D. mayri* and *D. birchii* plus the previously analyzed *D. biarmipes* and *D. virilis* CTR sequences.

Dotplot analyses comparing CTR sequences of DINE-TR1 from *D. biarmipes* and *D.virilis* with SatDNA-7 consensus from *D. serrata*, *D. mayri* and *D. birchii* showed that are all DINE-TR1 CTR-related (Figure 9A). The highest interspecific nucleotide identity (~87.5%) between these DINE-TR1 CTR concentrates in the first 30 bp of the sequences (Figure 9B), a feature also observed by Dias *et al*., (2015).

In order to investigate if these DINE-TR1 CTR sequences are expanded in *D. serrata*, *D. mayri* and *D. birchii* (as previously found in *D. virilis*, *D. americana* and *D. biarmipes*), we conducted MEGABLAST searches on NCBI using each SatDNA-7 consensus sequence retrieved for each species. The results are shown in Table SM3. Checking the top 10 contigs (sorted by total score) retrieved for each species, we found the CTR sequences expanded in satDNA-like arrays in all three species. In *D. serrata*, due to the availability of the long contigs, we were able to detect uninterrupted arrays up to ~82.6kb (~540 tandem copies). In *D. mayri* and *D. birchii*, the longest arrays have ~4.6kb (32 tandem copies) and ~1.6kb (11 tandem copies), respectively. Therefore, our findings confirm that SatDNA-7 are derived from DINE-TR1 CTRs that have expanded the copy number to long satDNA-like arrays in *D. serrata*, *D. mayri* and *D. birchii*. Accordingly, SatDNA-7 genomic proportion is relatively high in *D. mayri* (2.5%) and *D. serrata* (2.0%) (Figure 7). These figures are close to the genomic proportion of expanded DINE-TR1 CTRs found in *D. virilis* (1.6%) and *D. americana* (2.2%) (Silva *et al*., 2019).

In summary, our results with the SatDNA-7 showed a third independent event where CTRs from a DINE-TR1 gave rise to satDNA arrays. This reinforces the importance of DINE-TR1 as a potential source for the emergence of satDNA repeats, as suggested by Dias *et al*., (2015).

Concluding remarks

In the present work, we studied for the first time the satDNA landscape in *Drosophila* species from the *montium* group. For this purpose, we used the TAREAN pipeline for a *de novo* identification of satellite DNAs in the genome of 23 species with recently sequenced genomes (Bronski *et al*., 2020). We identified 142 satDNAs which correspond to 101 satDNA families present in different subgroups. The satDNA families are composed of different monomer sequences, which varies from 4 bp to >1,000 bp and present genomic proportions from 0.015% (SatDNA-2 from *D. auraria*) to 16.0% (SatDNA-18 from *D. pectinifera*). Probably, this expressive genome proportions variation is associated to gain or loss of satDNAs in heterochromatic regions of the chromosomes in different species, as suggested by Baimai (1980) and Venkat and Ranganath (2007).

Although the sequence size of satDNAs is widely variable, we found a prevalence of short satDNAs (≤10 bp) in the *montium* group. We also found that the 101 satDNA families present in the *montium* group are in most cases (76.45%) AT rich. This was expected because high AT richness is a remarkable feature of satellite sequences found in the genomes of different eukaryotes, including *Drosophila* (Schmidt, 1980; Ganal and Hemleben, 1986; Palomeque and Lorite, 2008; Melters *et al*., 2013).

Regarding the correlation tests, we found a strong negative correlation between the monomer size and the copy number of satDNAs. It is possible that this is a result of selection limiting the size of the genomic region(s) occupied by satellite DNAs. For example, if the optimal size occupied by a

satDNA array is around 1,000 bp, the number of satDNA copies will be constrained by the monomer size, so that this region could be filled by either 100 copies of 10 bp or 10 copies of 100 bp.

On the other hand, our correlation tests between genome sizes (predicted by Bronski *et al*., 2020) and genome proportions of tandem repeats and satDNAs were statistically non-significant. These results may suggest that other repetitive sequences (such as TEs, for example) may play a major role for genome size differences across species, or that our satDNA filter was insufficient to recover all satDNA sequences from the genomic data analyzed.

From the 101 satDNAs found in the *montium* group, only 17 were found in more than one species. The vast majority (83%) of the satDNAs are therefore species-specific. Among the 17 satDNA families, 5 are shared by 2 species, 3 by 3 species, 4 by 4 species, 3 by 6 species, 1 by 7 species and 1 by 13 species. These results are in accordance to the very fast rate of evolution usually reported for satDNAs (Ugarković and Plohl, 2002; Plohl *et al*., 2012).

We also tested whether satDNAs can be useful markers to infer phylogenetic relationships between species from the *montium* group. From the 17 satDNA families shared by at least two species, we studied their distribution across the montium group phylogeny, copy number and topology of ML trees. We concluded that SatDNA-1, SatDNA-2 and SatDNA-3 are useful markers for the *montium* subgroup, SatDNA-4 is a useful marker for the *punjabiensis* subgroup, SatDNA-5, SatDNA-6, SatDNA-7 and SatDNA-8 are useful markers for the *serrata* subgroup, SatDNA-10 and SatDNA-12 are useful markers for the *kikkawai* subgroup and SatDNA-11 is a useful marker for the *kikkawai/serrata* clade. The distribution of the 17 satDNA families are mostly in accordance with the phylogenies proposed by Yassin (2018) and Conner *et al*., (2021) at a subgroup level. Nonetheless, our data did not permit to infer additional phylogenetic relationships among the subgroups since we did not find satDNA families shared by all (or almost all) analyzed species.

Lastly, we found 12 satDNA families sharing homology to *Drosophila* Helitrons TEs. In particular, the SatDNA-7 satDNA family is an outcome of an expansion of the central tandem repeats (CTRs) present in DINE-TR1 (a Helitron element) in the *serrata* subgroup. This finding highlights the DINE transposable elements as a source for satDNA emergence, as previously showed in other *Drosophila* species (Dias *et al*., 2015).

Acknowledgments

The authors would like to thank Rafaella Soares, Pedro Heringer and Matheus de Morais for their valuable comments and suggestions.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S.,

Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., … Craig Venter, J. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, *287*(5461), 2185–2195. https://doi.org/10.1126/science.287.5461.2185.

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379.

Allen, S. L., Delaney, E. K., Kopp, A., & Chenoweth, S. F. (2017). Single-molecule sequencing of the *Drosophila serrata* genome. *G3: Genes, Genomes, Genetics*, *7*(3), 781–788. https://doi.org/10.1534/g3.116.037598.

Ashburner, M., Bodmer, M., & Lemeunier, F. (1983). On the evolutionary relationships of *Drosophila melanogaster*. *Developmental genetics*, 4(4), 295-312.

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., Schubert, I., & Macas, J. (2018). Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-24196-3.

Baimai, V. (1980). Metaphase karyotypes of certain species of the *Drosophila montium* subgroup. *The Japanese Journal of Genetics*, 55(3), 165-175.

Bock I. R., Wheeler M.R. (1972). The *Drosophila melanogaster* species group. *Studies in genetics* VII. Univ Texas Publ 7213:1–102.

Bosco, G., Campbell, P., Leiva-Neto, J. T., & Markow, T. A. (2007). Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics*, *177*(3), 1277–1290. https://doi.org/10.1534/genetics.107.075069

Bronski, M. J., Martinez, C. C., Weld, H. A., & Eisen, M. B. (2020). Whole genome sequences of 23 species from the *Drosophila montium* species group (Diptera: Drosophilidae): A resource for testing evolutionary hypotheses. *G3: Genes, Genomes, Genetics*, *10*(5), 1443–1455. https://doi.org/10.1534/g3.119.400959.

Conner, W. R., Delaney, E. K., Bronski, M. J., Ginsberg, P. S., Wheeler, T. B., Richardson, K. M., Peckenpaugh, B., Kim, K. J., Watada, M., Hoffmann, A. A., Eisen, M. B., Kopp, A., Cooper, B. S., & Turelli, M. (2021). A phylogeny for the *Drosophila montium* species group: A model clade for comparative analyses. *Molecular Phylogenetics and Evolution*, *158*(April 2020), 107061. https://doi.org/10.1016/j.ympev.2020.107061.

Craddock, E. M., Gall, J. G., & Jonas, M. (2016). Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. *Genetica*, 144(1), 107-124.

Da Lage, J. L., Kergoat, G. J., Maczkowiak, F., Silvain, J. F., Cariou, M. L., & Lachaise, D.

(2007). A phylogeny of *Drosophilidae* using the Amyrel gene: Questioning the *Drosophila melanogaster* species group boundaries. *Journal of Zoological Systematics and Evolutionary Research*, *45*(1), 47–63. https://doi.org/10.1111/j.1439-0469.2006.00389.x.

Da Silva, M. J., Fogarin Destro, R., Gazoni, T., Narimatsu, H., Pereira Dos Santos, P. S., Haddad, C. F. B., & Parise-Maltempi, P. P. (2020). Great Abundance of Satellite DNA in *Proceratophrys* (Anura, Odontophrynidae) Revealed by Genome Sequencing. *Cytogenetic and Genome Research*, *160*(3), 141–147. https://doi.org/10.1159/000506531.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, *7*(12). https://doi.org/10.1371/journal.pgen.1002384.

de Lima, L. G., Svartman, M., & Kuhn, G. C. S. (2017). Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. *G3: Genes, Genomes, Genetics*, *7*(8), 2831–2843. https://doi.org/10.1534/g3.117.042093.

Dias, G. B., Svartman, M., Delprat, A., Ruiz, A., & Kuhn, G. C. S. (2014). Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biology and Evolution*, *6*(6), 1302–1313. https://doi.org/10.1093/gbe/evu108.

Dias, G. B., Heringer, P., Svartman, M., & Kuhn, G. C. S. (2015). Helitrons shaping the genomic architecture of *Drosophila*: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. *Chromosome Research*, *23*(3), 597–613. https://doi.org/10.1007/s10577-015-9480-x.

Dias, C. A. R., Kuhn, G., Svartman, M., Santos, J. E. D., Santos, F. R., Pinto, C. M., & Perini, F. A. (2021). Identification and characterization of repetitive DNA in the genus *Didelphis* Linnaeus, 1758 (Didelphimorphia, Didelphidae) and the use of satellite DNAs as phylogenetic markers. *Genetics and molecular biology*, 44.

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 1–19. https://doi.org/10.1186/1471-2105-5-113.

Feschotte, C., & Wessler, S. R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 98(16), 8923-8924.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet*., 41, 331-368.

Gall, J. G., Cohen, E. H., & Polan, M. L. (1971). Repetitive DNA sequences in *Drosophila. Chromosoma*, *33*(3), 319–344. https://doi.org/10.1007/BF00284948.

Gall, J. G., & Atherton, D. D. (1974). Satellite DNA sequences in *Drosophila virilis*. *Journal of Molecular Biology*, *85*(4), 633–664. https://doi.org/10.1016/0022-2836(74)90321-0.

Ganal, M., & Hemleben, V. (1986). Different AT-rich satellite DNAs in *Cucurbita pepo* and

Cucurbita maxima. *Theoretical and applied genetics*, 73(1), 129-135.

García, G., Ríos, N., & Gutiérrez, V. (2015). Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus *Austrolebias* (Cyprinodontiformes, Rivulidae). *Genetica*, *143*(3), 353–360. https://doi.org/10.1007/s10709-015-9834-5.

Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. *Genes*, *8*(9). https://doi.org/10.3390/genes8090230.

Garrigan, D., Kingan, S. B., Geneva, A. J., Andolfatto, P., Clark, A. G., Thornton, K. R., & Presgraves, D. C. (2012). Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Research*, *22*(8), 1499–1511. https://doi.org/10.1101/gr.130922.111.

Han, M. V., & Zmasek, C. M. (2009). PhyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, *10*, 1–6. https://doi.org/10.1186/1471-2105-10-356.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), 5–9. https://doi.org/10.1093/nar/gkn201.

Kapitonov, V. V., & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15), 8714-8719.

Kim, B. Y., Wang, J., Miller, D. E., Barmina, O., Delaney, E. K., Thompson, A., ... & Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *eLife*, 10, e66405.

Kohany, O., Gentles, A. J., Hankus, L., & Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics*, 7(1), 1-7.

Kuhn, G. C. S. (2015). Satellite DNA transcripts have diverse biological roles in *Drosophila*. *Heredity*, *115*(1), 1–2. https://doi.org/10.1038/hdy.2015.12.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096.

Lohe, A. R., Hillikert, A. J., & Roberts, P. A. (1993). Mapping Simple Repeated. *Genetics Society of America*.

López-Flores, I., & Garrido-Ramos, M. A. (2012). The repetitive DNA content of eukaryotic genomes. *Genome Dynamics*, *7*, 1–28. https://doi.org/10.1159/000337118.

Melters, D. P. (2013). Comparative Analysis of Tandem Repeats from Eukaryotic Genomes: Insight in Centromere Evolution. University of California, Davis.

Miller, D. E., Staber, C., Zeitlinger, J., & Hawley, R. S. (2018). Highly contiguous genome assemblies of 15 *Drosophila* species generated using nanopore sequencing. *G3: Genes, Genomes, Genetics*, *8*(10), 3131–3141. https://doi.org/10.1534/g3.118.200160.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & MacAs, J. (2013). RepeatExplorer: A

Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792–793. https://doi.org/10.1093/bioinformatics/btt054.

Novák, P., Robledillo, L. Á., Koblížková, A., Vrbová, I., Neumann, P., & Macas, J. (2017). TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, *45*(12). https://doi.org/10.1093/nar/gkx257.

Palomeque, T., & Lorite, P. (2008). Satellite DNA in insects: A review. *Heredity*, *100*(6), 564–573. https://doi.org/10.1038/hdy.2008.24.

Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA evolution. *Repetitive DNA*, 7, 126-152.

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, *6*(June), 1–14. https://doi.org/10.1038/srep28333.

Russo, C. A., Mello, B., Frazão, A., & Voloch, C. M. (2013). Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zoological Journal of the Linnean Society*, 169(4), 765-775.

Schmidt, E. R. (1980). Two AT-rich satellite DNAs in the chironomid *Glyptotendipes barbipes* (Staeger). *Chromosoma*, *79*(3), 315-328.

Sena, R. S., Heringer, P., Valeri, M. P., Pereira, V. S., Kuhn, G. C. S., & Svartman, M. (2020). Identification and characterization of satellite DNAs in two-toed sloths of the genus *Choloepus* (Megalonychidae, Xenarthra). *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-76199-8.

Silva, B. S. M. L., Heringer, P., Dias, G. B., Svartman, M., & Kuhn, G. C. S. (2019). *De novo* identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines. *PLoS ONE*, *14*(12), 1–15. https://doi.org/10.1371/journal.pone.0223466.

Sullivan, L. L., Chew, K., & Sullivan, B. A. (2017). α satellite DNA variation and function of the human centromere. *Nucleus*, *8*(4), 331–339. https://doi.org/10.1080/19491034.2017.1308989

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, *9*(4), 678–687. https://doi.org/10.1093/oxfordjournals.molbev.a040752.

Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences. *Exs*, *67*, 21–28. https://doi.org/10.1007/978-3-0348-8583-6_2

Thomas, J., & Pritham, E. J. (2015). Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiology spectrum*, 3(4), 3-4.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing:

Computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117.

Ugarković, Đ., & Plohl, M. (2002). Variation in satellite DNA profiles—causes and effects. *The EMBO journal*, 21(22), 5955-5959.

Venkat, S., & Ranganath, R. A. (2007). Localization and characterization of heterochromatin among four Species of the *montium* subgroup of *Drosophila*. *Cytologia*, 72(3), 279-286.

Yassin, A. (2018). Phylogenetic biogeography and classification of the *Drosophila montium* species group (Diptera: Drosophilidae). *Annales de La Societe Entomologique de France*, *54*(2), 167–175. https://doi.org/10.1080/00379271.2018.1447853.

Yunis, J. J., & Yasmineh, W. G. (1971). Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science (New York, N.Y.)*, *174*(4015), 1200–1209. https://doi.org/10.1126/science.174.4015.1200.

Supplementary Table 1. General features for each satDNA sequences retrieved from our TAREAN analyses.

SatDNA Family	Species	Consensus size (in bp)	Proportion	Number of reads	Base pairs	Copy number estimate	AT content [%]	Satellite confidence	Satellite probability	C value	P value	Similarity hits – TAREAN database [above 0.1]	*Drosophila* Repbase megahits
1	*D. asahinai*	374	0.200	511	433,955	1,160	74.34	Low	0.6870	0.935	0.872	No	DNA4-1_DK#RC/Helitron
1	*D. rufa*	368	0.160	437	337,230	916	72.29	Low	0.3720	0.920	0.828	No	DNA4-1_DK#RC/Helitron
1	*D. lacteicornis*	374	0.260	558	529,037	1,414	74.07	High	0.775	0.962	0.931	No	DNA4-1_DK#RC/Helitron
1	*D. tani*	374	0.570	1786	1,110,475	2,969	74.07	High	0.727	0.951	0.853	No	DNA4-1_DK#RC/Helitron
1	*D. auraria*	367	4.600	22639	10,121,660	27,579	75.48	High	0.993	0.986	0.958	No	DNA4-1_DK#RC/Helitron
1	*D. triauraria*	366	5.400	20453	11,719,986	32,021	75.41	High	0.994	0.989	0.966	No	DNA4-1_DK#RC/Helitron
2	*D. asahinai*	163	6.300	16174	13,669,610	83,862	61.97	High	0.992	0.995	0.984	No	No
2	*D. rufa*	163	5.400	14448	11,381,540	69,825	63.20	High	0.985	0.997	0.979	No	No
2	*D. lacteicornis*	163	6.600	14249	13,429,407	82,389	61.35	High	0.986	0.992	0.987	No	No
2	*D. tani*	188	4.500	14169	8,766,908	46,632	66.49	High	0.976	0.994	0.942	No	No
2	*D. auraria*	166	0.021	104	46,207	278	68.68	Low	0.07010	0.846	0.857	No	No
2	*D. triauraria*	165	0.015	58	32.555	197	68.49	High	0.751	0.948	0.871	No	No
3	*D. asahinai*	17	0.360	938	781,120	45,948	76.48	High	0.993	0.989	0.962	No	No
3	*D. rufa*	17	0.150	405	316,153	18,597	76.48	High	0.976	0.988	0.938	No	No
3	*D. lacteicornis*	17	0.310	668	630,775	37,104	76.48	High	0.981	0.998	0.965	No	No
4	*D. punjabiensis*	20	0.920	7688	1,816,522	90,826	95.0	High	0.853	0.996	0.931	No	No
4	*D. watanabei*	20	0.490	2436	892,779	44,638	85.0	Low	0.66100	0.988	0.860	No	No
5	*D. mayri*	190	0.920	1326	2,055,265	10,817	72.64	High	0.889	0.956	0.900	No	Rehavkus-1_DY#DNA/MULE-NOF ; MuDR-

													1_DK#DNA/MULE-NOF ; FB4_DM#DNA/TcMar-Tc1 ; PARISa_Dan#DNA/TcMar-Tc1
5	*D. birchii*	121	0.140	592	236,808	1,957	77.69	Low	0.0276	0.747	0.721	No	Rehavkus-1_DY#DNA/MULE-NOF ; MuDR-1_DK#DNA/MULE-NOF ; FB4_DM#DNA/TcMar-Tc1 ; PARISa_Dan#DNA/TcMar-Tc1
6	*D. serrata*	30	1.400	6492	2,585,434	86,181	56.67	High	0.959	1.000	0.998	No	No
6	*D. bunnanda*	66	2.700	8125	4,893,753	163,125	54.55	High	0.959	0.999	0.999	No	No
7	*D. mayri*	150	2.500	3637	5,584,960	37,233	70.0	High	0.780	0.972	0.882	No	Helitron-2N1_Dvi#RC/Helitron ; Helitron-N2_DBi#RC/Helitron ;
7	*D. serrata*	153	2.000	9382	3,693,477	24,140	75.17	Low	0.00935	0.856	0.524	No	Helitron-2N_DW#RC/Helitron
8	*D. truncata*	198	4.100	14767	7,818,219	39,485	67.18	High	0.989	0.989	0.946	No	DNA4-1_DK#RC/Helitron
8	*D. truncata*	194*	1.200	4280	2,288,259	11,795	73.72	Low	0.64000	0.976	0.845	No	DNA4-1_DK#RC/Helitron
8	*D. birchii*	182	0.210	895	355,212	1,951	75.28	High	0.993	0.987	0.963	No	No
9	*D. birchii*	24	0.130	567	219,893	36,648	83.34	High	0.979	0.996	0.989	No	No
9	*D. bunnanda*	24	0.130	387	235,625	39,270	83.34	High	0.974	0.979	0.985	No	No
10	*D. bocki*	41	0.410	3918	635,891	15,509	65.86	High	0.985	0.999	0.981	0.33 organelle/plastid	No
10	*D. leontia*	41	0.890	5648	1,449,977	35,365	65.86	High	0.959	0.999	0.997	0.35 organelle/plastid	No
11	*D. bocki*	109	0.380	3636	589,363	5,407	73.40	High	0.986	0.995	0.973	No	No
11	*D. leontia*	109	0.660	4177	1,075,264	9,864	73.40	Low	0.0861	0.788	0.990	No	No
12	*D. bocki*	489	0.160	1498	248,152	507	30.48	High	0.981	0.983	0.958	No	No

12	*D. leontia*	489	0.100	655	162,918	333	30.48	High	0.926	0.959	0.961	No	No
13	*D. seguyi*	16	4.500	9322	9,306,656	1,163,332	87.5	High	0.985	0.999	0.985	No	Gypsy15-LTR_Dya#LTR/Gypsy; Gypsy3-I_Dpse#LTR/Gypsy
13	*D. vulcana*	16	5.500	9556	11,505,307	1,438,163	87.5	High	0.979	1.000	0.990	No	Gypsy15-LTR_Dya#LTR/Gypsy; Gypsy3-I_Dpse#LTR/Gypsy
13	*D. mayri*	8	0.300	437	670,195	83,774	87.50	High	0.992	0.993	0.977	No	No
13	*D. bakoue*	16	0.210	2367	460,546	57,568	87.5	High	0.994	0.988	0.966	No	No
13	*D. pectinifera*	8	0.450	900	990,985	123,873	75.0	High	0.959	1.000	0.996	No	No
14	*D. nikananu*	191	2.700	12166	5,878,088	30,775	71.21	High	0.956	0.982	0.907	No	No
14	*D. nikananu*	207*	0.990	4406	2,155,299	10,412	68.60	High	0.944	0.987	0.933	No	No
14	*D. nikananu*	188	2.400	10836	5,224,967	27,792	67.03	Low	0.02790	0.939	0.672	No	DNA4-1_DK#RC/Helitron; SAR2_DM#Satellite; SAR_DM#Satellite
14	*D. jambulina*	208	5.300	8010	9,511,839	45,729	69.24	Low	0.0489	0.821	0.763	No	DNA4-1_DK#RC/Helitron; SAR2_DM#Satellite
14	*D. seguyi*	178	1.100	2183	2,274,960	12,780	69.67	High	0.789	0.967	0.888	No	SAR2_DM#Satellite; DNA4-1_DK#RC/Helitron
15	*D. jambulina*	18	0.260	394	466,618	5,846	88.89	High	0.986	1.000	0.970	No	No
15	*D. bunnanda*	18	0.130	405	235,625	26,180	88.89	High	0.979	0.995	0.985	No	No
15	*D. serrata*	18	0.040	191	73,869	8,207	88.89	High	0.944	0.995	0.910	No	No
15	*D, seguyi*	27	0.280	581	579,080	64,342	88.89	High	0.950	0.983	0.917	No	No
15	*D. seguyi*	7	0.340	704	703,169	100,452	85.72	High	0.986	1.000	0.972	No	No
16	*D. mayri*	16	2.000	2892	4,467,968	558,496	87.5	High	0.979	0.999	0.992	No	No
16	*D.*	16	0.092	766	181,652	22,706	87.5	High	0.908	0.960	0.925	No	No

	punjabiensis												
16	*D. bakoue*	16	0.170	1855	372,823	46,602	87.5	High	0.959	0.999	0.997	No	No
16	*D. bocki*	16	0.026	243	40,324	5,040	87.5	Low	0.6970	0.975	0.855	No	No
16	*D. birchii*	8	0.100	448	169,148	21,143	87.5	High	0.979	0.996	0.991	No	No
16	*D. pectinifera*	10	0.930	1845	2,048,037	204,803	90.0	High	0.985	0.999	0.975	No	No
17	*D. bakoue*	18	0.210	2343	460,546	51,171	88.89	High	0.975	0.985	0.951	No	No
17	*D. punjabiensis*	18	0.061	507	120,443	13,382	88.89	Low	0.562	0.963	0.824	No	No
18	*D. pectinifera*	175	16.000	30914	35,235,045	20,.343	65.15	High	0.767	0.983	0.893	No	No
19	*D. birchii*	150	3.200	14022	5,412,759	36,085	78.67	High	0.992	0.995	0.985	No	No
20	*D. bunnanda*	346	3.100	9346	5,618,753	16,239	76.59	High	0.981	0.980	0.964	No	DNA4-1_DK#RC#Helitron; Helitron-N1_DEI#RC#Helitron
21	*D. serrata*	270	0.460	2160	849,499	3,146	71.12	High	0.986	0.996	0.967	No	No
22	*D. bunnanda*	324	5.900	17839	10,693,757	33,005	70.99	Low	0.0215	0.899	0.615	No	DNA4-1_DK#RC/Helitron ; Helitron-N1_DEI#RC/Helitron ; Helitron-N2_DT#RC/Helitron ; hAT-N1_DF#DNA/hAT?
22	*D. serrata*	356	10.000	49012	18,467	51,874	68.83	Low	0.1890	0.933	0.786	No	DNA4-1_DK#RC/Helitron ; Helitron-N1_DEI#RC/Helitron
23	*D. kanapiae*	31	0.400	7624	621,960	20,063	51.62	High	0.956	0.979	0.912	No	No
24	*D. bunnanda*	22	2.100	6233	3,806,252	346,022	36.37	High	0.981	0.976	0.958	No	No
25	*D. seguyi*	384	1.600	3325	3,309,033	8,617	49.22	High	0.981	0.976	0.956	No	Gypsy-12_DRh-LTR#LTR/Gypsy
26	*D. nikananu*	328	0.760	3365	1,654,572	5,044	59.76	High	0.922	0.976	0.926	No	No
27	*D. truncata*	92	0.960	3446	1,830,607	19,897	55.44	High	0.993	0.987	0.962	No	No
28	*D. pectinifera*	658	0.350	698	770,766	1,171	52.59	High	0.923	0.968	0.977	No	No
29	*D. bunnanda*	17	0.130	395	235,625	13,860	52.95	High	0.986	0.987	0.995	No	No
30	*D. kanapiae*	800	0.610	11489	948,489	1,185	65.0	Low	0.562	0.978	0.834	No	No

31	*D. burlai*	46	0.370	1513	733,079	15,936	52.18	Low	0.161	0.902	0.841	No	No
32	*D. bunnanda*	581	0.130	387	235,625	405	66.27	High	0.986	0.992	0.995	No	No
33	*D. pectinifera*	24	0.140	275	308,306	12,846	75.0	High	0.952	0.982	0.993	No	No
34	*D. jambulina*	10	5.900	8800	10,588,651	1,058,865	60.0	High	0.959	1.000	1.000	No	No
35	*D. triauraria*	4	0.260	983	564,295	141,073	100.0	High	0.756	0.980	0.883	No	No
36	*D. watanabei*	20	0.730	3613	1,330,059	266,011	80.0	High	0.979	0.998	0.990	No	No
37	*D. pectinifera*	26	1.500	2916	3,303,285	412,910	76.93	High	0.979	0.999	0.992	No	No
38	*D. birchii*	18	0.380	1657	642,765	71,418	88.89	High	0.959	0.997	0.996	No	No
39	*D. serrata*	18	0.150	692	277,010	46,168	83.34	High	0.985	0.996	0.983	No	No
40	*D. bakoue*	16	0.350	3969	767,578	95,947	87.5	High	0.979	0.998	0.987	No	No
41	*D. bocki*	202	0.400	3806	620,382	3,071	67.33	High	0.992	0.993	0.980	No	DNA4-1_DK#RC/Helitron
42	*D. vulcana*	28	0.100	178	209,187	29,883	57.15	High	0.992	0.989	0.978	No	No
43	*D. watanabei*	81	0.160	778	291,519	32,391	81.49	High	0.892	0.974	0.950	No	Gypsy-13_DSim-I#LTR/Gypsy
44	*D. birchii*	10	0.360	1581	608,935	60,893	80.0	High	0.994	0.994	0.974	0.38 organelle/mitochondria	No
45	*D. truncata*	6	1.100	3972	2,097,571	349,595	83.34	High	0.985	0.997	0.976	No	No
46	*D. jambulina*	21	0.200	306	358,937	17,092	76.2	High	0.959	1.000	1.000	No	No
47	*D. burlai*	8	0.900	3643	1,783,167	222,895	87.5	High	0.986	0.996	0.971	No	No
48	*D. bakoue*	4	0.320	3607	701,785	175,446	0.0	High	0.895	0.963	0.954	No	No
49	*D. tani*	216	0.190	592	370,158	1,713	34.73	High	0.901	0.975	0.897	No	No
50	*D. jambulina*	21	0.160	246	287,149	41,021	85.72	High	0.968	0.988	1.000	No	No
51	*D. burlai*	18	0.800	3258	1,585,037	176,115	77.78	High	0.955	0.982	0.943	0.18 organelle/plastid	No
52	*D. bakoue*	7	0.220	2417	482,477	68,925	85.72	High	0.979	0.999	0.989	No	No
53	*D. seguyi*	9	0.810	1684	1,675,198	186,133	88.89	High	0.985	0.999	0.977	No	No
54	*D. pectinifera*	12	0.410	813	902,898	75,241	75.0	High	0.979	1.000	0.993	No	No
55	*D. truncata*	26	0.100	372	190,688	14,668	80.77	High	0.979	0.997	0.989	No	No
56	*D. bunnanda*	15	0.230	679	416,875	27,791	60.0	High	0.979	1.000	0.991	No	No
57	*D. pectinifera*	16	0.390	771	858,854	107,356	62.5	High	0.985	1.000	0.982	No	No
58	*D. mayri*	18	0.450	644	1,005,292	55,849	88.89	High	0.922	0.981	0.934	No	No

59	*D. bunnanda*	20	0.220	651	398,750	19,937	65.0	High	0.979	1.000	0.991	No	No
60	*D. bocki*	21	0.110	1026	170,605	8,124	95.24	High	0.989	0.993	0.954	No	No
61	*D. burlai*	325	0.370	1500	733,079	2,255	55.70	High	0.992	0.993	0.979	No	No
62	*D. mayri*	16	0.310	447	692,535	86,566	87.5	High	0.979	1.000	0.987	No	No
63	*D. burlai*	1897	0.320	1306	634,015	334	61.26	High	0.944	0.992	0.935	No	No
64	*D. seguyi*	7	0.330	686	682,488	97,498	85.72	High	0.985	0.997	0.983	No	No
65	*D. pectinifera*	18	0.290	567	638,635	70,959	66.67	High	0.981	0.998	0.962	No	No
66	*D. bakoue*	16	0.150	1721	328,962	41,120	75.0	High	0.979	0.999	0.992	No	No
67	*D. bunnanda*	174	0.140	425	253,750	1,458	69.55	High	0.851	0.993	0.897	No	DNA4-1_DK#RC/Helitron ; SAR2_DM#Satellite ; SAR_DM#Satellite
68	*D. seguyi*	9	0.290	602	599,762	66,640	55.56	High	0.932	0.963	0.974	No	No
69	*D. mayri*	16	0.220	321	491,476	61,434	87.5	High	0.985	1.000	0.981	No	No
70	*D. mayri*	22	0.170	240	379,777	17,262	63.64	High	0.986	1.000	0.967	No	No
71	*D. bakoue*	8	0.120	1354	263,169	32,896	62.5	High	0.959	0.999	0.997	No	No
72	*D. mayri*	285	0.150	212	335,097	1,175	51.93	High	0.848	0.943	0.945	No	No
73	*D. bakoue*	14	0.100	1136	219,308	31,329	85.72	High	0.994	0.986	0.972	No	No
74	*D. mayri*	8	0.140	201	312,757	39,094	87.5	High	0.965	1.000	0.951	No	No
75	*D. seguyi*	16	0.170	357	351,584	43,948	81.25	High	0.981	1.000	0.962	No	No
76	*D. mayri*	49	0.120	170	268,078	5,470	46.94	High	0.755	0.947	0.932	No	No
77	*D. seguyi*	40	0.140	295	289,540	7,238	42.5	High	0.780	0.966	0.879	No	Gypsey-12_DRh-LTR#LTR/Gypsy
78	*D. seguyi*	16	0.130	263	268,858	33,607	87.5	High	0.985	1.000	0.977	No	No
79	*D. bocki*	186	0.380	3629	589,363	3,168	69.9	High	0.922	0.983	0.934	No	DNA4-1_DK#RC/Helitron
80	*D. mayri*	422	0.870	1257	1,943,566	4,605	71.33	High	0.975	0.978	0.946	No	Jockey-8_Dvi#LINE/I-82Jockey ; Merlin-1_DF#DNA/Merlin
81	*D. leontia*	653	0.310	1975	505,048	773	48.86	Low	0.5830	0.945	0.827	No	DNA4-1_DK#RC/Helitron
82	*D. nikananu*	239	1.200	5325	2,612,483	10,930	62.35	High	0.986	0.997	0.969	No	(GAAA)n#Simple_Repeat
83	*D. seguyi*	264	0.180	370	372,266	1,410	71.97	High	0.757	0.941	0.897	MuDR-_DT#DNA/MULE-NOF; MuDR-	Rehavkus-1_DY#DNA/MULE-NOF ; MuDR-1_DK#DNA/MUL

												_DK#DNA/MULE-NOF; MuDR-_DF#DNA/MULE-NOF; FB4_DM#DNA/TcMar-Tc1	E-NOF ; FB4_DM#DNA/TcMar-Tc1
84	*D. mayri*	189	9.500	13675	21,222,850	112,290	70.9	High	0.986	0.993	0.989	No	DNA4-1_DK#RC/Helitron
85	*D. burlai*	135	3.100	12458	6,142,020	45,496	70.38	High	0.986	0.996	0.974	No	No
86	*D. kanapiae*	59	0.770	14517	1,1972,74	20,292	66.11	High	0.950	0.979	0.916	No	No
87	*D. pectinifera*	22	1.700	3320	3,743,723	170,169	72.73	High	0.979	1.000	0.993	No	No
88	*D. birchii*	10	0.450	1960	761,169	76,116	70.0	High	0.985	0.998	0.982	No	No
89	*D. jambulina*	423	0.380	576	681,980	1,612	51.78	High	0.981	0.976	0.973	No	No
90	*D. vulcana*	27	0.270	463	564,806	20,918	55.56	High	0.989	0.989	0.954	No	No
91	*D. punjabiensis*	455	0.520	4371	1,026,730	2,256	62.64	High	0.989	0.990	0.952	No	DNA4-1_DK#RC/Helitron
92	*D. burlai*	441	1.700	6895	3,368,204	7,637	45.58	High	0.992	0.991	0.977	No	No
93	*D. tani*	138	0.650	2020	1,266,331	9,176	65.22	High	0.993	0.995	0.955	No	No
94	*D. kanapiae*	18	0.150	2805	233,235	25,915	88.89	High	0.793	0.916	0.944	No	No
95	*D. punjabiensis*	601	0.330	2785	651,578	1,084	42.60	High	0.955	0.976	0.941	No	No
96	*D. punjabiensis*	30	0.150	1268	296,172	9,872	70.0	High	0.992	0.990	0.975	3.08 organelle/mitochondria	No
97	*D. bunnanda*	19	0.350	1039	634,375	33,388	94.74	High	0.974	0.982	0.983	No	No
98	*D. bunnanda*	20	0.110	316	199,375	19,937	50.0	Low	0.4260	0.911	0.915	No	No
99	*D. truncata*	10	1.300	4766	2,478,947	247,894	80.0	High	0.965	0.995	0.948	0.13 organelle/plastid	No
100	*D. seguyi*	10	0.710	1470	1,468,383	146,838	70.0	High	0.994	0.993	0.971	No	No
101	*D. bunnanda*	19	0.190	586	344,375	18,125	73.69	High	0.986	0.986	0.986	No	No

* Other consensus sequence for the same satDNA in one species

Supplementary material 2. Satellite DNA consensus sequences identified with TAREAN in species from the *montium* group.

>374_D.asahinai_LC_SatDNA-1
AATCTAGTATAAAAACCTTCAAAACTATGCCAAAACAGCCTTAATTTCAATTTGGAAAGCTTTTCTG
TGTCAGTTTTTATTAGATTAATAAATCTGCTTATTAAGATTAGAAATTTAAACAAATAATTTTTTTGC
GATGTTTTGAAAAAACTAAGTTTCCCCCTTATCAAATTTTGAAAAATCGATAAAAAATTGTTTTCTGA
TATTTTTCAATATTTTGGTTTTAAGTGTTTACAATTTTAACGACGAATCGGAAAATACACGGCTTTT
GGCTATCACTTCCTTTTTGGCTATAAGAGGCTCTTAAGAAGACCTTACTTTCTCACATTTTTTGAAT
CAAAAAATGTTTCGAAAAATATGATTTTTTTTTTAAT
>368_D.rufa_LC_SatDNA-1
GGGAAATCTAAGATAAAAACCTTCATAACTAAGCCAAAACAGCCTTAATTTCAATTCGGAAAGCTT
TTCTGTGTCAGTTTTTATTAGATTAATAAATCTGCTTATTAAGATTAGAAATTTAAGCAAATAAATTT
TTCGTGATTTTTTGAAAAAATACTGCCATCCCCCCTTATCACATTTTGAAAAATCGATAAAAAAAAT
TTTTCTGATATTTTTCAATATTTTGGTTTTAAGTGTTTACAATTTTAACGACGAATCGGAAAATACAC
GGCTTTTGGCTATCACTTCCTTTTTGGCTATAAGAGGCTCTTAAGAAGACCGTACTTTTGATTCAA
AAAATGTGTCGAAAAATGTTTTTTTTTTTTTTT
>374_D.lacteicornis_HC_SatDNA-1
CAAAAAATGTGGAAAAGTAAAATCTTCCTAAGAGCCTCTTATAGCCAAAAAGGAAGTGATAGCCAA
AAGCCGTGTATGTTCCGATTCGTCGTTAAAATTGTAAACACTTAAAACCAAAATATTGAAAAATATC
AGAAAACAATTTTTTATCGATTTTTCAAAATTTGATAAGGGGGAAACTTAGTTTTTTCAAAAAATCG
CAAAAAAATTATTTGTTTAAATTTCTAATCTTAATAAGCAGATTTATTAATCTAATAAAAACTGACAC
AGAAAAGCTTTCCGAATTGAAATTAAGGCTGTTTTGGCATAGTTTTGAAGGTTTTTATCCTAGATTA
TTAAAAAAAAAATCATATTTTTCGAAACATTTTTTGATT
>374_D.tani_HC_SatDNA-1
TTAAAAATAAAATTCCAAATATATGATTTTTTTTGAAAAATCTAAGATAAAAATTACCATAACTAAGT
CAAAACAGCCTTAAATTCAATTCGGGAAGCTGTTCTATGTTAGATTTTATTAGCTTATAAATTCTGC
TTATTTAGTTTAGAATTCTTGACAATTAAATTTTTTGCCAATTTTCGACATTTTTAATGATGAAAGTT
ATGAAAATTTTGAAAAATCGATCAAAATTTTTTTTCTGGAATTTTTTAATATTTTGGCTTTAAGTGTTT
ACAATTTTAACGACGAATCCGAAAATACACGGCTTTTGGCTATCACTCGCTTTTTGGCTATAAGAG
GCTCTTAAGAAGACCTCAAATTTTCACATTTTTGGG
>367_D.auraria_HC_SatDNA-1
AAACTGAGTATGCAGATTTATTATTTTCATAAAACCTAACATAGAACAGCTTTCCAAATTGAAATTA
AGGCTGTTTTAGCTTAGTTATGACTATTTTTGTCTAAGTTTTTTTCAAAAAATTATCATATTTAAAAA
ATAAAAATATTTTGAAAATTTAAGGTCTTCTTCAGCTCCTCCTATAGGCAAAAAGGGAGTGATAGC
AAAAAGGCGTGTATTTTCGGATTCGTCATTAAAATTACAAACATTTAAAACCAAAATATTGAAAAAT
ATCAGAAAAAATTTTTTTTTTTGATTTTTTAAAATGTTAGAAGGTCGGGGGTCGTTAAAATGTTCAAA
AATTGGCATAAAATTTATTTGTTTAAATTTTT
>366_D.triauraria_HC_SatDNA-1
TAACGACCCCCGACCTTCTAACATTTTAAAAAATCAAAAAAAAAAATTTTCTGATATTTTTCAATATT

TTGGTTTTAAATGTTTGTAATTTTAATGACGAATCCGAAAATACACGCCTTTTTGCTATCACTCCCT
TTTTGCCTATAGGAGGAGCTGAAGAAGACCTTAAATTTTCAAAATATTTTTATTTTTTAAATATGATA
ATTTTTTGAAAAAAACTTAGACAAAAATAGTCATAACTAAGCTAAAACAGCCTTAATTTCAATTTGG
AAAGCTGTTCTATGTTAGGTTTTATGAAAATAATAAATCTGCATACTCAGTTTAAAAATTTAAACAAA
TAAATTTTATGCCAATTTTTGAACATTT

>163_D.asahinai_HC_SatDNA-2

GTGTTATTCTGCCCATAACTTCTAAACGACTTAAGCTACAGAAATGATGTAAGTACTGTTGTCTTCT
GGCGTTAAATACGCAGCCAACGACACCACATTCATCCCGATCGGATGCTCCGTTGAAAAGATATT
CAATAAAGATGATTTTACCGTTGTTTTTTTT

>163_D.rufa_HC_SatDNA-2

AACGGTAAAATCATCTTTATTGAATATCTTTTCAACGGAGCATCAGATCGGGATGAATGTGGTATC
GTTGGCTGCGTATTTAACGCCAGAAGACAACAGTACTTACATCATTTCTGTAGCTTAAGTCGTTTA
GAAGTTATGGGCAGAATAACACAAAAAAAAC

>163_D.lacteicornis_HC_SatDNA-2

GATTTTACCGTGGTTTTTTTTGTGTTATTCTGCCCATAACTTCTAAACGACTTAAGCTACAGAAATG
ATGTAAGTACTGTTGTCTTCTGGCGTTAAATACGCAGCCAACGACACCACATTCATCCCGATCGG
ATGCTCCGTTGAAAAGATATTCAATAAAGAT

>188_D.tani_HC_SatDNA-2

ATATTTTTTTAACTATCTATATCTATTCTTTATTGAATATCTTTTCAACGGAGCATCAGATCGGGATG
AATGTGGTGTCGTTGGCTGCGTATTTAACGGCAGAAGACAACAGTACTTACATTATTTCTGTAGCT
TAAGCCGTTTAGAAGTTATGGGCAGAATAACACAAAAAAAACATCGAAAAAATC

>166_D.auraria_LC_SatDNA-2

AAGTTTTTACAGTTTAATTTTTTTTAATTACTTCTGCCCATAACTTCTAAACGGCTTAAGTTACAGAA
ACAGTTTAAATGTTGTTTTGCTCTAGTTTTGAGTACGCGTCGAACAATACCTCAATCATCCCGATC
AGATACTCTGTTAAAAAAAATACTCAACAAAG

>165_D.triauraria_HC_SatDNA-2

TACTTCTGCCCATAACTTCTAAACGGCTTAAGTTACAGAAACAGTTTAAATGTTGTTTTGCTCTAGT
TTTGAGTACGCGTCGAACAATACCTCAATCATCTCGATCAGATACTCTGTTGAAAAGATATTCAAC
AAAGAAGTTTTTACAGTTTAATTTTTTTTAAT

>17_D.asahinai_HC_SatDNA-3

ATGCATGAATTTTTTTC

>17_D.rufa_HC_SatDNA-3

TGCATGAATTTTTTTCA

>17_D.lacteicornis_HC_SatDNA-3

CATCGATGAAAAAAATT

>20_D.punjabiensis_HC_SatDNA-4

AAAAATATATCAAAATATAT

>20_D.watanabei_LC_SatDNA-4

CATGCAAAAATATATAAAAA

>190_D.mayri_HC_SatDNA-5

TTTTTAGAGTTTCTGGATGGAATTTTTGAGTTTTAAGGGGTTTTAAGGGGGCAAACGTTTGTTATTT
TGTTTTAAGAGTTATTTGATGAAATTGTTTAGTTTTAAGGGGTGGGCATTTTAGATTTTGTAGTTTT
AAGAGATGGGCAGAAGTTTTTTTCTTAAATTTTTTAGTTTTAAGGATTTTTTTTAT
>121_D.birchii_LC_SatDNA-5
TTAGTTTTAAAGGGTGGGCATTTTATATTTTGTAGTTTTAAGAGATGGGCAGAAGTTTTTTTTTTTA
ATTTTTTAGTTTTAAGGATTTTTTTTATTTTTTAGTATTTCTCGATGAAAATGT
>30_D.serrata_HC_SatDNA-6
GTCAGAAGTGTCAGTTTCAGATGTGTCAGT
>66_D.bunnanda_HC_SatDNA-6
CTACTGACACTGACACTTCTGTCACTGACACTTTTGGCACTGACACTTCTGACACTGACACTTTCA
>150_D.mayri_HC_SatDNA-7
ATATTATTCCTGAATTTCGGTTTTAATTTCATGAAAATCGGCCTACTATATCCTATAGCTTCCATAG
GAACAATCAGGAAAATAAAAGAAAACCCCATAATTTTGTCAGTTTTCAACATATTTTCATCAAAGTT
GAGATTTAGCCGTTTT
>153_D.serrata_LC_SatDNA-7
TTTCTCGATCGTTCCTATGGGAGCTATATGATATAGTAGTCCGATTTTCATAAAATTTTACACAAAA
TTGTAGAGTAATATAAAATGGCTATATCTCAAATTTGATGAAATTCTGTTGAAAACTGAAAAAGTTA
TAATTTTTAATTGTATTAA
>198_D.truncata_HC_SatDNA-8
CCTTATCTTTTTTTGGAATAATGGCAACATGAAAATCGAAGTTTGGGAATGCCATAACTTGGCCAA
AAATCCATCAATTTCAATTCTGGAAGCACAGAAGTGTTTGTTTTTATTAGCTTAACAACACTGCAAA
CCAAATTTATTAAATTTCCATAGTTTCAGATTTTTCGAATTTTTCACAATTTTGACGACCCCCAC
>194_D.truncata_LC_SatDNA-8
ACATCAACAAAAAAATCGAAGTTGGGAATACCATAAATTGGCCAAAAATCTATCAATTTTAAATCTG
AGAGCACAACAGTATTATTTTTATTAGCTTAACAAACCAGCAAACCCAATTAAATAATTTTCATTAAT
ATCAAAGATTTTGAATTTTTCACAATTTTCACGACCCCTTATCATTTTCGAAAAAAAAA
>182_D.birchii_HC_SatDNA-8
TTAAAAAATTCAAATTTTTTAATTTAAATAAATTGATAAATTTAGGTTTGCAGCTTTGTAGAGCTAAC
AAAAACAAGCACTTCTGTGCTTACAGAATTAAAATTAGTTTATTTAAAGCCAAGTTATGGCATTTCA
AAATTTTGATTTTCATGATTTATTTTTTAAGATAAGGGGGTACCCCA
>24_D.birchii_HC_SatDNA-9
GAATAAGAATAAGAATAAGAATAA
>24_D.bunnanda_HC_SatDNA-9
TTATTCTTATTCTTATTCTTATTC
>41_D.bocki_HC_SatDNA-10
CCTTCTTTTTTGCCTTCTTCTTCGTTATTTAGTCTTCTACT
>41_D.leontia_HC_SatDNA-10
GAAGAAGGCAAAAAAGAAGGAGTAGAAGACTAAATAACGAA
>109_D.bocki_HC_SatDNA-11
AGGCTAATACATTTGCATACTCAATTGTAAATTTGTAAACAAAAAGTTCACAACTCAGCTAATAATG

CATGTATCCACTTCAGAATTTGTATTTATATTAGTTTTTACA
>109_D.leontia_LC_SatDNA-11
TTTACAAATTTACAATTGAGTATGCAAATGTATTAGCCATGTAAAAACTAATATAAATACAAATTCTG
AAGTGGATACATGCATTATTAGCTGAGTTGTGAACTTTTTG
>489_D.bocki_HC_SatDNA-12
AGCCGCCCTCCGCTGCCGACCCCCGGCGAATACGCCGCCGCCGCTGCCCGACGGAAAGCCGA
GGAGCTGTGCGACGAGCTGTGCTCGTCATCCACCGAGCCTCCCCAGTTGGCCGGTCCGGAGCA
ACACTCGGCCCGTTATCCACCACCGACGCCGGCCCCTGCTCGACGAATCCGCCACGATGGCAG
AAGACCGCTCGTCCCAGCACGTCCACGCCGCCGCTTTATCAGCGAGTGGCCCCGGTCCATGGA
CGACCTCTTCGGGGATTCACCGGACCAGGACGGCGCGGAGTTGCCGACGAGATCTTCGGGCCG
GGCCCCAGATACGCTGCGCTGCAGCCGAGGTGTCACCGGCAGCAGCACCTAGCCGCCCGGAG
GTTACGCAGCCGTCGTCGCCCGACGCCGACCTGAAGGGCCGTCGACGAGCCGTCCTCCCTTAT
CGACGACCGACGAGGACTCTGCAGCCGAAATGCCGCTGGCAGCAGCACCA
>489_D.leontia_HC_SatDNA-12
GGTGCTGCTGCCAGCGGCATTTCGGCTGCAGAGTCCTCGTCGGTCGTCGATAAGGGAGGACGG
CTCGTCGACGGCCCTTCAGGTCGGCGTCGGGCGACGACGGCTGCGTAACCTCCGGGCGGCTAG
GTGCTGCTGCCGGTGACACCTCGGCTGCAGCGCAGCGTATCTGGGGCCCGGCCCGAAGATCTC
CTCGGCAACTCCGCGCCGTCCTGGTCCGGTGAATCCCCGAAGAGGTCGTCCATGGACCGGGGC
CACTCGCTGATAAAGCGGCGGCGTGGACGTGCTGGGACGAGCGGTCTTCTGCCATCGTGGCGG
ATTCGTCGAGCAGGGGCCGGCGTCGGTGGTCGATTACGGGCCGAGTGTTGCTCCGGACCGGCC
AACTGGGGAGGCTCGGTGGATGACGAGCACAGCTCGTCGCACAGCTCCTCGGCTTTCCGTCGG
GCAGCGGCGGCGGCGTATTCGCCGGGGGTCGGCAGCGGAGGGCGGCTT
>16_D.seguyi_HC_SatDNA-13
TATTTAGTTATTTAGT
>16_D.vulcana_HC_SatDNA-13
TATTTAGTTATTTAGT
>8_D.mayri_HC_SatDNA-13
TAACTAAA
>16_D.bakoue_HC_SatDNA-13
CTAAATAACTAAATAA
>8_D.pectinifera_HC_SatDNA-13
CTAGTTAT
>191_D.nikananu_HC_SatDNA-14
GAATGCAGAAATGTTCTCCTATAAAAACTAAGAAGAAAAAACGAAACCAGGCTGAAATCGGTTGA
GTTTAAGCCAAGTTATGATTAAAAAATTAAAGAAAAGGGGAGCGATTTTTAAGGCACTAAAAAAAA
GAAAAGGGGTTTCATCATAATAATTTTTCAAAATTGACAAAAAATTGGAAATCAAGTTTT
>207_D.nikananu_HC_SatDNA-14
GCTAAGAAGAAAAAACGAATCCAGACTCAAATCGGTTGAGTTTAAGCCAAGTTATGATCAAAATCC
CTTGGAGGTTTTAGCGATTTTTAGGCCAAAAATATGACCTATTTAAAAAACAAATTGCAAGGGGTT
TTCTCATCATTTTTTACAAAACTTGGCAAAAAATGAGAAATCGAGTTTTAAATGCAGAAAAGTTCTT

CTATTAAA
>188_D.nikananu_LC_SatDNA-14
TTTTGGCTCAAAAATCGCGAACTTCTTCTGTGGTTTTTTGATCGTAACTTGGCTTAAACTCAACCG
ATTTCAGTCTGGTTTGGTTTTTTCTTCTTGGTTTTTAATGGAGAACATTTCTGCATTTAAAACTCGAT
TTCTAATTTTTGGCCAATTTTTGAGAATTTTGACGATGTAACCCCTTTAATTTT
>208_D.jambulina_LC_SatDNA-14
TTTAGTATTTTTCAGCCAAAAAATTTGACCCAATTTTTGAAAAAAATGTAAGGGGTTACATCGTTAA
AATTGTCAAAAATTGGCCAAAAATTAGAAATCGAGTTTTAAATGCAGAAATGTTCGCATTTAAAAAC
CAAGAAGAAAACAGCAAACCAGACTGAAATCGGTTGAGTTTTAGCCAAGTTATGATCATAAAACCA
ATGAGGGT
>178_D.seguyi_HC_SatDNA-14
CATGGTTTTTTGATCATAACTTGGCTTAAACTCAACCGATTTCAGTCCGGTTTGGTTTTTTCTTCTT
GGTTTTCAAAGGAGAACATTTCTGCATTCAAAAATTTAATTTCTAAATTTTTCCCAATTTTTTTAAAT
TTTGACGATGTAACCCCTTAAATTTTGTTTGTCCCAAAAATCG
>18_D.jambulina_HC_SatDNA-15
ATATATGAAATATATGAA
>18_D.bunnanda_HC_SatDNA-15
TATGAAATATATGAAATA
>18_D.serrata_HC_SatDNA-15
TATGAAATATATGAAATA
>27_D.seguyi_HC_SatDNA-15
TTCATATATTTCATATATTTCATATAT
>7_D.seguyi_HC_SatDNA-15
TGAAATA
>16_D.mayri_HC_SatDNA-16
TGTATATTTGTATATT
>16_D.punjabiensis_HC_SatDNA-16
TGTATATTTGTATATT
>16_D.bakoue_HC_SatDNA-16
ATATACAAATATACAA
>16_D.bocki_LC_SatDNA-16
ACAAATATACAAATAT
>8_D.birchii_HC_SatDNA-16
ATACAAAT
>10_D.pectinifera_HC_SatDNA-16
TATAACTATA
>18_D.bakoue_HC_SatDNA-17
TTATTATTGTTATTATTG
>18_D.punjabiensis_LC_SatDNA-17
ACAATAATAACAATAATA

>175_D.pectinifera_HC_SatDNA-18
TAAATCGGCATTTTTATATCCTAATTTTTGGACAAATCAAACTTAAGAATAACCATAACTTGGCCAA
AACAGCCTTAAAAACAGCGATTCGGAATATATATGGCATTTGGCTGTCATTCCATTTTTGGCCAAA
TTATGCTCTTCAGAAGACATCAAATGCGGATTTTTGGGTCCT
>150_D.birchii_HC_SatDNA-19
CATTTATTAGCCTATTTCATGAAATAGACCAATAATTCAAATTATTACACGAAATATTTTTTGATTAT
TATTGTTAACAAAAAATGTGTCTAGTTATTGAGGAAAATGTAATTTAATACATAAAACAGGTTTAAAT
TACACAAATAAACC
>346_D.bunnanda_HC_SatDNA-20
AATTTTAACGATGAGATCCCTTAAAAAACTTGAAAAAAATTTTAAAAATTTATTTTTCTTTTAAATATA
GCTAGATCGATTCGCCTGTTGATTCTGATTAAGAATATATATGGATTACAGGGTCGGAAATAACTC
CTTCACAGCGTTACAAGCTTCTGGGTTAAATTACAATATCATTCTTGATCGATATTTATCTATAATTT
TATTGAACAATTGTTCAAAAAATCAAAAAATAAAAATCTCGGAATTCGGAAAGCTTTTCTATGTTAG
TTTTTCTTAGGTTAATAATTCTGCATACTAAATTTCAAATTTTTAAAAATTTAAATTTTTATCAATTTTT
TTTTTT
>270_D.serrata_HC_SatDNA-21
GTAAAACTAAATATTTTACCGAGTTAAGCATACTTAATAAAACAAAACATCATTTGTACCGCGTTTT
AAATATATATTTTGCGTTCGTCCAGCTTCTTAAATAAAATAGTTTATTTAAAGTAGCACCAAGTAATA
CTGTTGGGGCGCCCCTTAAGCATAAATTATATTGTTTTCCCAACCACGTATTCAATTTATAAAAATA
AACAATGCCCAACTGAATAATATGAATATAACAACCCTCAATAATAAATCAAATTGCTTCTAAACCG
A
>324_D.bunnanda_LC_SatDNA-22
ACATTTTTCGATATTTTTATATGAGTATTTTTTCGACAAATCAAGATTTAAAAATCTCATAACTAAGC
CAAAAAAGCATTAATTTTAATTCGGAAAGCTGTTCTATGCTAGTTTTTAGTAGGTTAATAAATCTGC
ATACTCAGTTTAAAACTTAAAAAAATTCAATTTTTGGCCAATTTTTGAAAATTTTAGTTTTCGTCCGG
AGATGGCTAGATCGATTCGGCTGTTGATGCTGATCAAGAATATATATGATTTATAGGGTCGGAAAT
GACTCCTTCACTGCGTTACAAGCTTCTGAATAAAAGTATAATACCATTTTTTAGT
>356_D.serrata_LC_SatDNA-22
GAGGAAAATATAATTTTTATCAGTGAAGGAGCCATTTCCGACCCTATAAATCATATATATTCTTGAT
CAGCATCAACAGGCGAATCTTTCTGGCCATCTCAGGTTGCAAAACAGACATTTTCAATTTTTTCTA
AAAAATTAAAGGGGTTACATCATTAAAATTTGCAAAAAAAGGCCAAATCTTAACAAGTTAAATGGA
GTATGCAGATTTGTTAAGCGCCAAACAACTAACACAGAACCGCTTTCCGAATTGAAATTAATGCAT
TTTTGGCAAAGTTATGAGATTTTTAAACTTTGACTTTCCCAAAAAATACTAAAACAATTATGAAATAA
AAATTTCGGTTTCAGGAATCGTA
>31_D.kanapiae_HC_SatDNA-23
TGGTCTGATGACCTGATCTGGTAAACAGACC
>22_D.bunnanda_HC_SatDNA-24
CATGGCTGCTGCATGGCTGCTG
>384_D.seguyi_HC_SatDNA-25
CAAGTGTGACTTTTCATCCTCAATAGCCCATTTCCGGTCCCTTTCACGTGGCCATGGGCGGCAAT

GCGACGGGAGGCGTACGGTGCTCTCAACTGATCGGTCTCCTTCCTGACCATGTGTCTCACAGTG
GCTACTTCCTGTTGAAACTAGTGGTTACTAAAGATAAAGCAAATTTGTGGCTAAGTATATTTACCA
CATTGACGACTGCCTTGCGGGGCGCTGTCTAGATGGGAGCTGCCTAAGTAGGCGTACCTCACCC
GATCACCGTCGGACGAGGAGTGAAATTCCAAGGCTCTGGCCCGCCTTATATAGGGGCCTGACGA
TATTTAGTGTGCCCGCCGCCTTTAAGAAATTTGTCATATTAGAGTGGCCAACCCCATTTTT
>328_D.nikananu_HC_SatDNA-26
ATTATTAACACGGATGTGTACGGGTGAAAAGCTGCGGAGAGACTGAGGCTGCTGCGGCGAGTCG
CCTCTATTTATAGGTATGAAAGATGGAGAGATTATATAAATAAATAGATTTTTTGCGCTTACCGATC
GGTCCTTTCAGCTTTATATATAGGATATGTATATGATAATAGTCTTGGTTCGAACTTACCCGTATTA
GAGGACACTAGTTACCTTTCGTTGCAACACCGCAGACTTGGCCGAGATTCATCGATGATTAAGTC
CTTTTGAGTGCAGAGGTTCTTTACCCAGTTATTAACACACTTGTTTACACACTTTTTAACACGTG
>92_D.truncata_HC_SatDNA-27
TGATATGTACTTAAGCTCAGCTGCTAACTTAAGAACGACCATGGAAATCCTGGATTGTGACGTCAT
AAGATCGCCCTTCTCTGCGCTTCTCA
>658_D.pectinifera_HC_SatDNA-28
CAACCAGTATCCAGGAATCTCCTCCACCAACAATTAGGTGCCAACATTATCACTGTTTAGCATGAC
CCAGTTCAGGATACACATTTTCCCGCAAAACGAGTGGCCAATCCCGGCCAGAATAGAGCCAATCC
TCTCCAGAATCGCATATTGGGGAAAGATGAGCATAGCAATGCATCCAAAAATGTACAATTGCCAC
CCATCTTATTTGGAGCCACCTTAAGAGCCTGGGACTTGTCCTTTAGCATACACCTAGCTGCCACA
ATTTCACCGATTTAATGGGGGAGAGTCCATGAACAATCATTGCCAACGATATAGAGCCAATTATG
GACCAATCCTCTCCAGAATTCCACCTGATCCCGGCCAGAAATTGGCCAAGTATCAACAAACTCGC
CAGTTGCCGGTCTCATTTACAGCCCTCACTCCTCCATGCATGACCCTGGCCAATAAACATGACCC
GCCATGGGAGCAGTGGCCAATCCTTTCCAGAATTTACCACAATTTCCTCAGCATTTCTAGCAAAAT
TGGGAGAAATTACCATTGCATATTATAACCTTCAATCCATATCGGCGGAGGAGGGTCCATTATCC
CTCAGAAGCGACCAGTTCCAGCCATCCAAGGCCTAATCCTCTCCAGAAACAGCCGATCCCGGCC
AGAAATA
>17_D.bunnanda_HC_SatDNA-29
ATCACCGAGCTATCACA
>800_D.kanapiae_LC_SatDNA-30
ATTTAAGGCTTTTTACTGCCTTTGCACTTAAGCGCCTTTTAGTCGACAGCAGTTTTTAGTTTTTATT
TACAAATAATATGAAAGCTAGGTGCGCCAGCATTGTGTAAAATGTTTTTTACCTAAAAAGCTTTTCC
AATATGGTCATATAAATTGAAAAAGTAAAATATAGCTGCGCCATCTATAGTGTAATTTTTCTTACAA
TGTATATTTTAAGTAACGTCAAATGTGATTTAACTAATACTAACAGCTAGCTGCGCCATCTATTGTG
TAAAATGTGTGTATCATGTAAAATGTAGCAAGTACACATTATATTTCTTGACAATATGACAGTAGCG
TTTAGGTAGTGGTCTGGAGCTGATGTATACGCGTTCGGCTCCCGCGATCATACGACAAAATTGGT
CAACAACGTATAATGGACACATGTGTATTGTGACAAGTATATCATTGTGACATGGGATCAATGGCA
GAAGATAGGTGCAGTGGTAAATACTTTACCCACAGCATAGCGGGAAAGTGAAATAGTGGAGTGGT
GTGGTTTTAGAAGTTCCGCGGGTGTGGCCGAGTGGGAGATATGCCAAAATGTGCGCAACACAGT
CATTTTAAAGTATGCATGTGTGTTTTATACCAAATAAAGCTAGATAGTAAAAATATAAAATTATATTT
AATCTTTTTATTGCAGGTTGAGTATGCAACTTACAAAGTTTGTTTTCACGGTGGACAGTAGAGAAA

CAACAATAAAATATAGTGAAAATAAACGCAAATATGAGACAGCGGCAGCATTTTGGTAATATTTTTT
TACA
>46_D.burlai_LC_SatDNA-31
GACGCCATTTTGTCCCAATATTGCAACACAGCGATGGCGGTAGTAA
>581_D.bunnanda_HC_SatDNA-32
AAAAAGCGCCACTAATTCGGGGTTTTGAACTATTATGTGGCGCCAATTCGCGGGCACATACGGCA
ACTATACATTTTATACGTGTAAAACCTATTTTACTACGTAAATCAAAAAATATCAAAAACGCGCCAC
TAATATGGTTTTTTGAACTATTATGTGGCTCCATTTCGCGGCCGCATACGGTAAACATACAATTTAT
ACGTGTAGAGCATATTTTACCATAATTTTCCTAATTTTGCTAAAATTGCCAAAATTCAAAAATTCTCT
TATTATTAGCATTGCTCAGCGTCGTAAAGCTGATAAAACCGCTAGAAATCAAAAAATACATACAATT
TATACGTGTAAAACATATTTTCCTATGTATTTCTTAATTTTGCTAAAATGGCCAAAATCACAACATTT
ATATTATTATTAGCATTGCTCAGCGTCCTAAAGCTCATAAAACTGTTAGAAACACAAAAAATCGAAA
AAGCGCCACTTTTTGGTTTTTTGAACTATTACGTGGCGCCAAGTCGCGGCCGCATACGGCAACTA
TACAATTTACACGTGTAAAACGTATTTACTACGTAAATCAAAAAATC
>24_D.pectinifera_HC_SatDNA-33
ATACAAACACATGTAGAAAAACAT
>10_D.jambulina_HC_SatDNA-34
GAATAGAGGA
>4_D.triauraria_HC_SatDNA-35
TTAT
>20_D.watanabei_HC_SatDNA-36
TGATATGATATGATATGATA
>26_D.pectinifera_HC_SatDNA-37
TGTGATATTGTGATATATGTGATATT
>18_D.birchii_HC_SatDNA-38
AATTTCAAAAATTTCAAA
>18_D.serrata_HC_SatDNA-39
AGATATAGATATAGATAT
>16_D.bakoue_HC_SatDNA-40
TGATATATTGATATAT
>202_D.bocki_HC_SatDNA-41
CCCCCAATATTTCTTTAAAAAATTTTTAAAAATTTTGGGTAAATATAAAGGCAAAACTGAGTACGCA
GTTTTGTTGAGCGTAGAAATCCCAATTCAGAACAGCTTTCCGCATTGAATTTCATGCGTTTTTGGC
CGAGTTATGAATTTTCTAAAATGGTTATTTCCCCCGAAAAATTATCACAATAAATGGCACCCATTTT
TT
>28_D.vulcana_HC_SatDNA-42
GAAGGAACAAGGAACAAGGAAGAAGGAA
>81_D.watanabei_HC_SatDNA-43
GTTTGATTTGTTTGATTTGTTTGATTTATTTGATTTATTTGATTTATTTGATTTCTTTGATTTCTTTGA
TTTCTTTGATTT

>10_D.birchii_HC_SatDNA-44
GAAAAGATTT
>6_D.truncata_HC_SatDNA-45
TGAAAA
>21_D.jambulina_HC_SatDNA-46
TACTTAATACTTACTACTTAC
>8_D.burlai_HC_SatDNA-47
AATATATC
>4_D.bakoue_HC_SatDNA-48
CTGT
>216_D.tani_HC_SatDNA-49
GTTCCTTGGAGACGGCGATGACGATGATGTGGCCGTGGCGGCCCTTGGAGCCGATGGCGACGA
CGATGCGGCCGTGGTGGCCCTTGGAGACGGCAGCGACGATGATGTGGCCGTGGTGGCCCTTGG
CGACGATGACGATGGCGACGAGGATGCACCCGTGGTGGTTCTTGATGACGATGACGTGGCCCG
AGAGCCCGATGCTGTGGTGACCGGTCT
>21_D.jambulina_HC_SatDNA-50
ATAAATGATAAATGATAAATG
>18_D.burlai_HC_SatDNA-51
ACGAAAAATACGAAAAAT
>7_D.bakoue_HC_SatDNA-52
GTTTATT
>9_D.seguyi_HC_SatDNA-53
ACAATAAAT
>12_D.pectinifera_HC_SatDNA-54
TGTGATAGATAT
>26_D.truncata_HC_SatDNA-55
GTTCAAAAATTTTGTCCAAAAATTTT
>15_D.bunnanda_HC_SatDNA-56
CAACATCATCATCAG
>16_D.pectinifera_HC_SatDNA-57
AAGCATGAAAGCATGA
>18_D.mayri_HC_SatDNA-58
AAAATTTTGAAAAAATTG
>20_D.bunnanda_HC_SatDNA-59
TTACGTTACACTACACTACA
>21_D.bocki_HC_SatDNA-60
TTTTGAAAAAAATTTAAAAAA
>325_D.burlai_HC_SatDNA-61
ATTCCACCCCTTAGTGTGACCGTATTGGCAACTTTGTTAGGTGTGACCATATTGGCAAAAGTCGTA
TTTGTCTATCGATTAAGATCTGGCGTTGCCAGACTTTTCGAACGCAAGCCGATGGTGCTGCCAGA

CTTTTAACTTTTGAGGGATTTCGGCGTCTTTGTGTGCCCATCCGCCGTGGAGGCGGAAATGTTTT
TGTTGTATTGCCAATGTTGTTGTAGGCAAAAACAAGGTCTTAATACCGCCTTTTAGCAGTACAAAA
TCATCAAATGCTGGATTTCATGCACCAGGCGCGCAGAAGCAGCAGCAAGTAGATAATAAAAAA
>16_D.mayri_HC_SatDNA-62
GTTAAATAGTTAAATA
>1897_D.burlai_HC_SatDNA-63
TTCCCTACATGGGATTTGGACGTCGCCCAAATTGGCCAAAATTTGACTAAAAAATTGGGCCACGT
ATTTGAGCCTGGGACATCAACTAGATAAAGGTCCTAACTATTAGTATGCTTAAAACTATACTTGTAA
CTGCAAGGATGCCAAAGTTAAGTAAATTCACGGAAGGGTCTGCTGCTGATGGCTGGCGGCAGCG
CAGTCGGCGTCAATAAAAAATGACCTCTACCAAAAAGGGTCGAAATTGGCCTATAAATTCCTGCC
ACGATCTCGGACTTGATGGTGTCCACAAGAGGATGGTCTGAGCTACCTGTGTGCCTAGTTTGAGA
TATATAGCTGCATGGATGAAGGAGTTATGCGATCCTGCCAATTGACGATTTCTGACGAAATTTTTT
CCTCTCATCGTAATAATTCCAATCAAATCCGTTAAAATCCAAATGTAAATTGACGATGCGGAGAAA
TGTAAACAAAACGTGCACTTACCCTGGAAAAGAGAAGAAAATATTAATAAATATTCTAAAGACTGA
TTATATTTAAATTACTCACCTTTGCATATAGATGTAACCGGATGGTGATGATAAACAAACGTTGTTG
GCTGCAATTTCCGGGTATTATTGCCTGCTTGCCCTTTGGAAATAAAATATATCATTATTAGCTAAAT
GAATTAATTAAATGTTTTAACAAATATTTTAATAAATGTCAAATGTCTGTCATATAATTATCATAAAAT
TAACAATAACTTAAATAAATAATATAAATAATCTTAATTATTAAATAGCGACGTGGGATAAAACCAG
CAAGCAGTCAATAAATAAACAACAATAATAAAAAACCAACTCATTAATATCTTGCATACAAAACGAC
ATATTAAAAGATCCAATAAATTATTAAAAATATTATAAAAACTTAAATATTTAAAAATTAAAATATTTA
AAACGCATTTAAATTGGTAGACTTAGTACGGAGGTCGCCAAAAACTATCAAAAGGGGGATTAAAA
ACTTGGATACACATCTGGGGTCAATCATAGCCAATAGATAAAGGCCTTGGCTAGTTGGTTGCTTA
AAATGAGGTCCGTAACTCCAAGGATGCCGAAGGTATGCATTTGTTGCAATTGGAATCGATTATGC
TGGCTGTTGTAAGCGCCGTCCGCGCTAAGTTTTTGATGGTTTCCAAAAAAGGTCAAAACACAAAG
AATTTCTTGGACCACGCCCTTCTCACCTTTTGGGTTGCCATTAGTAAGGTCTAAGCTAACTATGTA
CACAAACTCAGGTCACTGGCTTCGAGGATGAGGGAGATCTCACTTTGCGTGAATTGTCAAAAACC
GCACATATGCACGATCCGTGCGTCAGTATGGAGGCCCCCAAAATCGGCGAAAATCTGACTTAAAC
ACGCTTTTTACGCTATTCTGACAAGATGGTCGCCATGAGGGAGGTTCAAGCTAGTTATGAACAAA
AAATTCAGGTCCCTAGCCCCAAGGACGAAGGAGATCTCACTTTGCGCAAATTGGGTAGACAATCA
CACAAATGCACCGTCCGTGCGTGTGTATGGAGGTCGCCAAAAATGGCCAAAAAAGGCTTAAAAAA
CCGTCCACGCAATCGGGCCTGGGGCTACTATTTGATAAAGGTCCAATCTAATTGTATGCAAAAAA
GTAGTCCTGTAACTGCAAGGATGCCTGTTATGACAATCTACAGGTCTTGCTGCCCGTATATGACG
GCAGTCTGTGCCGTAGTACAAAAATTACAAAAAATGGTCTTCGCCAAAAATGGCCAAAAAACGGA
CAAGTTTTTGGGCCACGCCTTCGAGCCTTTTTGGGTTGCCATTAGGATGGACTAAACTAACTGGC
TACATAAATTCAGACCACTAGCTCTAAGGACATTGGAGAAATCTTGACGCGTAGCA
>7_D.seguyi_HC_SatDNA-64
ACAAATA
>18_D.pectinifera_HC_SatDNA-65
ATATCACCAATATCACCA
>16_D.bakoue_HC_SatDNA-66

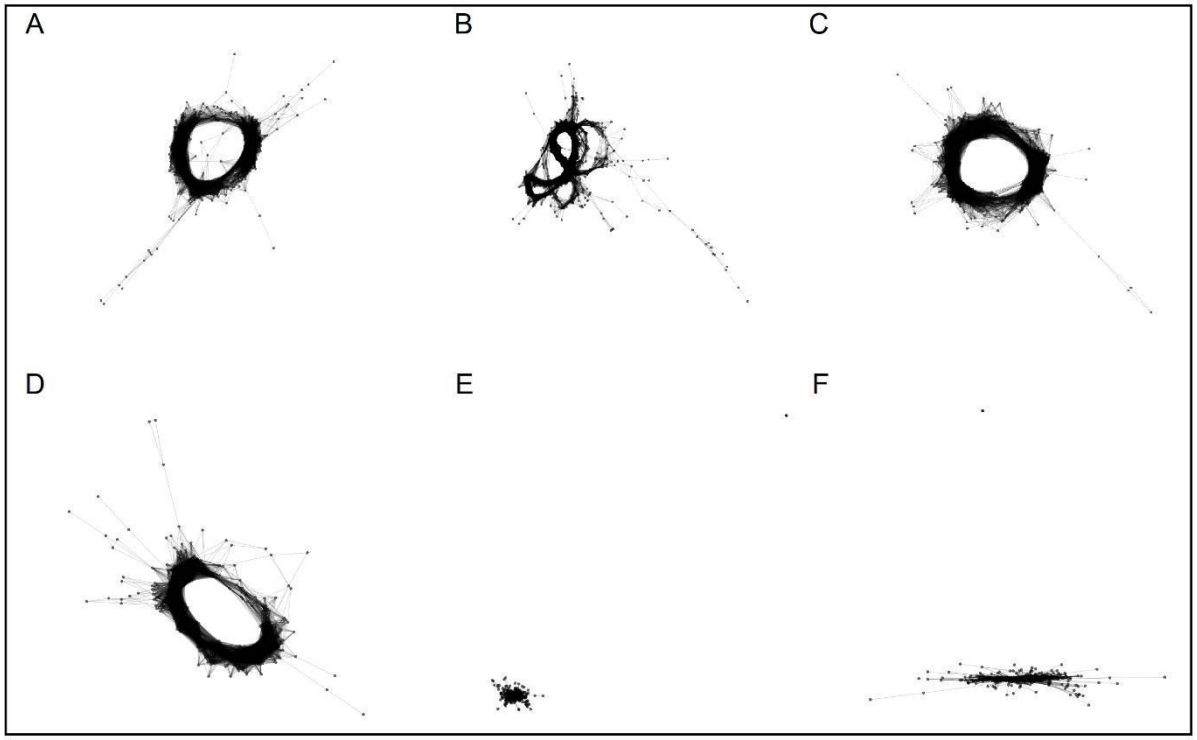
TCTTGATATCTTGTTA
>174_D.bunnanda_HC_SatDNA-67
TGAAATCGGTTGAGTTTAAGCAAAGTTATGATGAAAAAACTATTTCATATGAAGCTCGATTTTTTCA
TTTTTTGATAAGGGGTTACATCATTAAAATTGTCAAAAATTGGAAATCGAGATTTGAATGCAGAAAT
GATCTCCTATAAAAACCAAGAAGAACACACCAAACCGAAC
>9_D.seguyi_HC_SatDNA-68
CAACAACAG
>16_D.mayri_HC_SatDNA-69
AAGTATATAAGTTTAT
>22_D.mayri_HC_SatDNA-70
AAAACTGTCCAAAAACTATCGC
>8_D.bakoue_HC_SatDNA-71
TGATATGG
>285_D.mayri_HC_SatDNA-72
TGTTGCAGCCTGCCAATCTACGAAATTTGTGTAGCACAGAATTGGCGCTCTTCGACTGCACCTCC
GTCACAAATATTTTTCTTGTTGTTCTAGCAGTCAAATTCTGTGTCTCACTGCTTTACCGCTATATAG
TTGTTGCTGTTATTGCAGCCTACCTAATCTACGACCTCTGTGTTGCTCAGAATTGGCGGTCTTCGA
CTGCGACTCGCTCGCCGATGTCATGCTTGCTGTTCTAGCAGTCGACCTCTGTGTCGCACTGCTTT
AGCGCTCGCTTGTTGTTGCTGT
>14_D.bakoue_HC_SatDNA-73
TCTTTATTCTTTAT
>8_D.mayri_HC_SatDNA-74
ATGATAAA
>16_D.seguyi_HC_SatDNA-75
TATTTGTACATTTGTA
>49_D.mayri_HC_SatDNA-76
GTAGACTTGCGAGTCCGCCTCTCTCTGAAGACTTTCGAGTCCGCAAACT
>40_D.seguyi_HC_SatDNA-77
TGATTCTCTTGCTCTTGCGGCTGACCTAGCGGTCCCCAAG
>16_D.seguyi_HC_SatDNA-78
ATAAGTAAATAAGTAA
>186_D.bocki_HC_SatDNA-79
TACCTATAATTTATAAGAAATTGGCGGAACAACAATCGGCGGAAAATCTCATAGCTCAGCCAAAAA
TGCATGAAATCCAATTCGGAAAGCTGTTCTGAATAGGTTTTTCTAAGCTTAACAAAACTACATACTC
AGTTTTGCTAAAATACTTACCTATAATTTATAAAAAATATTTAACCTATTGTT
>422_D.mayri_HC_SatDNA-80
GTTTTAAAAATAAATCAAAATCAGCAGCAATTCAATTATCATAGTACACAGAGTCAGAGAAAGAGA
GTGCGGGAGAGCGCAGTCTGTTAGTTCAATGGGGTCATAAAATTGGTTTTAGCTGTGCGTTGATT
CTTAGTTTTTAATAAAGGGTTAGTTCAATGCTTGTGTATAATTTGAATGAATTATTATTTCGTCTTTA
GGTTCATGTATTTTTTTTTAATTTACCTAGAGCTTCAGGAAAACCACCATATGAATCCTTAAAGTAT

CTGGATTCCCTCGAAAAAATTTCTTATGGCTTATTGACTTAAATCAGTGTTTTTTACTGTCAGAAGA
AAATGGGGAAGTTTTACAGATTTTTCAAAAGCGATAAAAATAGTAATTTTTTAATTTATTGTTAAGTT
TAATTTTATTTTAAATTTTAA
>653_D.leontia_LC_SatDNA-81
TGAACCTGGCGCCTCCACCGGAGCCGGGCACCTCCGCCGGAGCCGGACGCCACCGCGGGAAC
CGGACGCCTCCACCGGATCCAGGCGCCACCGCGGGAAACGGGCAGGACCGCGTGAGCCGGGC
GCTTTCATCAATTTCTTATAAATTATAGGTAAACAATAGGTTAAAATTTTTAAATAAATTATAGGTAA
GTATTTTAGCAAAACTGAGTATTCAGAACAGCTTTCCGAATTGGATTTCATGCATTTTTGGCTGAG
CTATGAGATTTTCCGGCGATTTTTTCCCGCCATTTTTTTATAAATATAGGTAAAAAATCCGAGCGAC
CAAAAGTACCAGATCGTCAGAAATATATATCGGCAAAAAATTCCAGAAATTAGTTGGCTATGAAAA
TATCGATACTTGGTATTTCTCAAATATATTCTCTTGGTCACACTAAGCCTCGCCTCGGCCCTATAA
AAGACGGGCCACAGCAGTGGGGAAGCCATTCGCCTGTTGACCGGCGATCGGTCAAGTACTCCG
AGGGAGGAGTAACCAAGGAGTCATCCCGTGGAGTGGATCGCCGGATCAATGCGTCCATGTCCC
GGGCACCACCGCGGGAGCCGGGCGCTACCACCGGAGCCGGTCACCTCTGACGAAACCGGGCG
CCACCGCG
>239_D.nikananu_HC_SatDNA-82
TTTCTTTGCTCGTTTTCTTTTTTTCTTTTGTCTCTTCTAGTTCTCCCTCTTCCTTTTGTCTCTCTTTCT
TTCGCTCTCTTTCTTTTTTCTTTCTTCCTTTTGTCGCTCTTTCTTCTTTCTTTCTCTCCGTCTTCCTT
CTGTCTCTTCTTCTTTGGCTCCCTTTCTTCCTTTTGCCTTCCTTTCTTCTTTTTCTCTTCTTCCTTTT
GTCTCTCTTTTAGTTCTCTCTTCCTTTTTCTCTC
>264_D.seguyi_HC_SatDNA-83
TACTTTTTGATTTTCAAAAATGTTTGAAAATTTTCGAAGTTTCATATTTTCGAAACGGGACGTGGTA
AAAAAAAAATGAAAGAAGGGCAAACATGGTCAAACGAAGGGTAAAGATGGTCAACTTTCAAATTTC
AAAAATTCAAATTTTGATAGAACACTTTTGAATATTTTCTGAATTTGGAAAATTTTCAAAATTCTAAA
GGGTGGGCAAACGTGGTCAAACAATTCTAATGCGATTTCCAAAAAAAAATTTTGAAAAAAGTT
>189_D.mayri_HC_SatDNA-84
TTTTTTGGGAATTTCAATTATCATAACTTGGCCAACAGGAAACTAATTTTGAATCGGAAAGCTGTTC
TATGCTAGTTTTCAAAAGGCTAACAAATCTGCATACTTACTTTTTCAATAAACAAACGATTTAATTTT
TTTGAATTTTGACGATGCAACCCCTTACAAAATTTGAAAAATTTTTTTGGCTCA
>135_D.burlai_HC_SatDNA-85
AAAAAAAAAACGAACAAAAATTCAACTTTTGAAATCTGGCATTTAAGTATGCAATTTCATAGATACA
GAAAGAAATAGCACAGAAAAGCTGTCGGAATCAAAATTAAAGCATTTTTGACGGAATTGTAACTCC
TC
>59_D.kanapiae_HC_SatDNA-86
GACAAAATTATATCTAACATGCCATCAGAGATGCTCATCTAGCTATATTGCATTATTGG
>22_D.pectinifera_HC_SatDNA-87
TATCACCAATATATCACGAATA
>10_D.birchii_HC_SatDNA-88
AGATTTCACA
>423_D.jambulina_HC_SatDNA-89

TCGTTACAGGCGACGGTCACACTGAACTTCGCCAGGGCCCTATATAAGGCGGGCGACACTCTTG
GAATTTCACTCGGCGTCCGACAGCGATCCGATGAGGTAGCTTATATCCAGAGAACAACCCGCCA
GGCAGCTGCCAGTGTGGTGAATATACTAAGCAACAAATTTGCTTGATTTGTAGTAACTAATACTTA
CTCAACAGGAAGTAGCTACCCTGAGGCACATGTTCAGGATGAAGGCCGCAGGATTGAGAGCACC
CTGCACCTCCTGTCGCATAACCGCGCACAGCCCACAGTATAGGAAAAATAACATAAAGAATTCGA
GTCATAAGAAAATATTTGAATTTAATGCGCGAGCGCAATTGATCGTAAACGCCAGGTCACACTGG
ATCGAACAGGCGTTGGCCACACTAATACGGCAATT
>27_D.vulcana_HC_SatDNA-90
GTGAAAAGACTACCCATGATAGGGTTG
>455_D.punjabiensis_HC_SatDNA-91
GAAGCCTTTCTCTCTAACGGTCTGGCAACCCTGAAAATATGGATGAGAATCGCAGAGAAGGGCG
CTTCACCTTAACAGTATGTAATTTGCAGAGGAAGGCCTACCTTAAGTACACAGTATGAGAATCGCA
GAGAATTTGCGTTGCAAACTTCTGCCTGAAATTATAATACCCTGCAAGGGTATAAAAAGTGTAACC
CCGAGTTAAGTATAGAAAGCTAATATTTTTATGGCTGCCATATTTCATTTGATTAAAGTCTATCAAC
CGCCAAACATTAGATTAGTTGGTTTTAAACAAAGTTTGGAGTTTCTAGCCTTGACAACATCTATATG
TGAATATCTATGCAAAAAGAAAATTATACATAAAGAGGAAATGACAAGTGCAAAAAAGGGAATTGC
TTAGGTATGTAAATCGAAAAGAGAGGCACTTACCAACAGAGTATGTAGAATCGCAGAGG
>441_D.burlai_HC_SatDNA-92
TGGCGCGAGGACCGCCGTCTACCATTGACAATTTCTGGCCCAACATTGTGAGCGGATTGCATGG
GATCGCAGCAGAAGATCCAGCTCATCCAGCAAGTGGTGACGCCCAAGGGCGAGCTGACGAATGT
CCCGGTGAGTGTAAGCAGGGATATATTCAACCGAAAGTCAAACTAACGAGCAACTTCTTTCCCTC
AGATCGCCACCAATGCCAATGAAAAAAAATCTCGCACATGTGTGCATATAATTCTCAACTGGTTCT
AACCTCCCCGGTATGAAGACACGGCCAAGTCGCAGGTGAACAATGTGATTTAAGGAGATTGTGG
AGCGCGCCAAGCCGCTGTCGATCCGCCTGACTAACCCCAACGATCGTCTGAGTGGTGGAACTGA
GCTCGGCCGTCGCCCAGTCGTACGCCGCTGGCAAGGGCCAGGCTTCAACGCCGG
>138_D.tani_HC_SatDNA-93
CATCAGATAGGAATAAATGTGGTGTCGTTTGCTCCGTATTTAACACCAGAAGACAACAGTACCTAC
ATCGTTTCTGTTGCTTAAGCCTTTTAAAAGTTATGGGCAGATTATTTTTTATTGAATATATTTTGAAC
CGAG
>18_D.kanapiae_HC_SatDNA-94
TGAAAAATTTGAAAAATT
>601_D.punjabiensis_HC_SatDNA-95
CAAGAAGAAGAACAAGGAGAAATGGTTCCCCGCCCTCGTTGTGACGCCTACTACACAGGTGAGC
TGAGAATTTGTTTTCATACAGGATATAATAATGATTGATTGCTTTTCTACACAGGCCACAGTCCGC
ATCCGCGTGAAGGACGAGTACCTGGTACATTCGTCCAAGGACGGCCGTTACTATATGGTCCCGA
AGAAGGAGGCCACCGAGTACACCCGCGCAGTGGCAAGTTGCGTTCACCGCCTCCCAGTCAGCG
GTTGGAGTGGGAGCAGCAGCCCCCGGAGCAGTAGTTGCCTTGGTAATTACCGCTGTGTTGCCGC
CAACAGCGGGTGCAGGATCAGGCGCATCCGGAACGGGTTCGGCCACAGCAACGGCCACTACCT
CAGGTGGCGCTGCTGTCGTGGTAAGCTCGGCGGCCAGGAAGCAGGCCCTCAAGGCAAGCGCCA
TTCAGCATAGCCTAAAAGGACGACTGACGCCCTCGGCCGTGGCCAATCCAGTCAAGATGCACAC

GCCACGAGGAGCAGAAGCAGCGTCTGCCAAGGAGGTGGTCAACGAGAAGGAGAAGAATATCGG
CAAGGTGGTGTGCGTGGAGACAGAGTC
>30_D.punjabiensis_HC_SatDNA-96
CTTATTCTCAATCTCAATCTGAATCTCATT
>19_D.bunnanda_HC_SatDNA-97
TTAAAATTTAAATTTTTTG
>20_D.bunnanda_LC_SatDNA-98
GAATGAGGAGGAATGAGGAG
>10_D.truncata_HC_SatDNA-99
AAATATACCA
>10_D.seguyi_HC_SatDNA-100
GAAAGATGAT
>19_D.bunnanda_HC_SatDNA-101
ACAATGTAACAACACAATA

Supplementary material 3. Graph layouts of SatDNA-1 clusters retrieved by TAREAN. *A. D. asahinai* SatDNA-1 cluster. *B. D. tani* SatDNA-1 cluster. *C. D. lacteicornis* SatDNA-1 cluster. *D. D. rufa* SatDNA-1 cluster. *E. D. auraria* SatDNA-1 cluster. *F. D. triauraria* SatDNA-1 cluster.



Supplementary Table 2. SatDNA families in the *montium* group sharing homology with Helitron transposable elements.

SatDNA Family	Species	Consensus size (in bp)	Helitron Hit (s) (Repbase search)	SatDNA hit fragment (From/To) "bp"	Correspondent fragment on Helitron sequence (From/To) "bp"	Similarity value between 2 aligned fragments
1	*D. asahinai*	374	DNA4-1_DK#RC/Helitron	19-201	126-312	0.7845
1	*D. rufa*	368	DNA4-1_DK#RC/Helitron	23-212	117-312	0.7720
1	*D. lacteicornis*	374	DNA4-1_DK#RC/Helitron	134-316	126-312	0.7889
1	*D. tani*	374	DNA4-1_DK#RC/Helitron	50-230	137-320	0.7337
1	*D. auraria*	367	DNA4-1_DK#RC/Helitron	1-97	223-320	0.7959
1	*D. triauraria*	366	DNA4-1_DK#RC/Helitron	227-359	187-320	0.8000
7	*D. mayri*	150	Helitron-N4_DSer	25-150	435-562	0.7795
7	*D. serrata*	153	Helitron-N4_DSer	1-149	344-491	0.9195
8	*D. truncata*	198	DNA4-1_DK#RC/Helitron	49-191	174-315	0.7413
8	*D. truncata*	194	DNA4-1_DK#RC/Helitron	28-191	146-315	0.7305
14	*D. nikananu*	188	DNA4-1_DK#RC/Helitron;			
14	*D. jambulina*	208	DNA4-1_DK#RC/Helitron;	24-190	138-313	0.7193
14	*D. seguyi*	178	DNA4-1_DK#RC/Helitron	13-178	132-313	0.7229
20	*D. bunnanda*	346	DNA4-1_DK#RC#Helitron	1-175 229-340	10-184 188-299	0.7543 0.7857
22	*D. bunnanda*	324	DNA4-1_DK#RC/Helitron	48-314	10-322	0.8303
22	*D. serrata*	356	DNA4-1_DK#RC/Helitron	21-291	42-322	0.8168
41	*D. bocki*	202	DNA4-1_DK#RC/Helitron	54-148	223-320	0.7396
67	*D. bunnanda*	174	DNA4-1_DK#RC/Helitron	56-107	131-193	0.9259
79	*D. bocki*	186	DNA4-1_DK#RC/Helitron	46-138	223-315	0.7634
81	*D. leontia*	653	DNA4-1_DK#RC/Helitron	216-267	264-315	0.8269
84	*D. mayri*	189	DNA4-1_DK#RC/Helitron	19-177	145-314	0.7546
91	*D. punjabiensis*	455	Helitron-N4_DSer	138-186	1-49	0.9796

Supplementary Table 3. Top ten contigs (sorted by total score) containing copies of SatDNA-7 in *D. serrata*, *D. mayri* and *D. birchii*. This satDNA family is composed of expanded CTR sequences of the DINE-TR1/Helitron transposable element.

Species	NCBI Acession	Acession length (in bp)	CTR copies
Drosophila serrata	MTTC01000695.1*	82685	540
	MTTC01000697.1*	54943	359
	MTTC01001044.1	93247	428
	MTTC01000698.1*	44750	292
	MTTC01000455.1	32409	180
	MTTC01000722.1	44489	229
	MTTC01000425.1	48635	172
	MTTC01000872.1	59640	162
	MTTC01000440.1	155132	319
	MTTC01000910.1	164193	138
Drosophila mayri	VNJN01003419.1*	4663	32
	VNJN01006878.1*	4174	28
	VNJN01000679.1	5556	29
	VNJN01001897.1	3150	19
	VNJN01005652.1	6398	25
	VNJN01008327.1	3707	22
	VNJN01001002.1	3827	22
	VNJN01010393.1*	3074	21
	VNJN01002029.1*	2736	19
	VNJN01009599.1	3572	21
Drosophila birchii	VNKA01002716.1	5544	28
	VNKA01002522.1	2056	14
	VNKA01006541.1	2712	9
	VNKA01002635.1*	1526	9
	VNKA01003810.1*	1517	11
	VNKA01005003.1	2328	7
	VNKA01006941.1*	1528	10
	VNKA01000528.1*	1243	8
	VNKA01003310.1*	1253	9
	VNKA01004301.1	1719	8

* Contigs containing only CTR sequences

5. Conclusões

Em relação aos nossos objetivos:

“Analisar a capacidade dos pipelines RepeatExplorer e TAREAN na identificação dos DNAs satélites mais abundantes de *D. virilis* e *D. americana*.”

Os pipelines RepeatExplorer e TAREAN foram capazes de fazer a identificação *de novo* dos satDNAs abundantes de *D. virilis* e *D. americana*. Embora ambas pipelines sejam úteis e complementares entre si para a identificação de satDNAs, o TAREAN pode classificar clusters de satDNAs verdadeiros como satDNAs de baixa confiabilidade. Neste caso, sugerimos que seja feita uma análise separada e detalhada para cada cluster obtido, levando em consideração os valores gerados para C, P e probabilidade de satélite além do formato do grafo.

***“*Mapear os prováveis DNAs satélites mais abundantes identificados pelo RepeatExplorer e TAREAN em cromossomos metafásicos e politênicos de *D. virilis* e *D. americana*.”**

Nós mapeamos as seguintes famílias de satDNAs abundantes presentes em *D. virilis* e *D. americana*: Sat1, 172TR e 225TR em cromossomos metafásicos e politênico de *D. virilis* e *D. americana* e PvB370 em cromossomo politênico de *D. americana*. Os mapeamentos foram realizados por FISH e apresentaram resultados inéditos, como a localização das sequências da família 225TR em cromossomos metafásicos das duas espécies, além do mapeamento da família 172TR em cromossomos metafásicos e politênicos e a colocalização das famílias 172TR e PvB370 em cromossomo politênico de *D. americana*. Além disso, os mapeamentos das famílias de satDNA estão de acordo com os resultados obtidos para as análises *in silico* realizadas com os pipelines RepeatExplorer e TAREAN bem como confirmam o elevado grau de parentesco filogenético entre *D. virilis* e *D. americana*.

“Identificar com o pipeline TAREAN prováveis satDNAs nos 23 genomas recém-sequenciados de espécies do grupo montium”

Identificamos, pela primeira vez, sequências de satDNAs presentes nos genomas das 23 espécies do grupo *montium* sequenciadas. Após investigação dos clusters obtidos via TAREAN, identificamos 142 clusters de satDNAs no total, e 17 famílias de satélites

compartilhadas por, pelo menos, duas espécies. Nossos dados demonstram que os genomas das espécies analisadas são enriquecidos com satDNAs de sequência monomérica menor do que 10 pb (sequências simples).

“Avaliar se as famílias de satDNAs identificadas nas espécies do grupo *montium* são bons marcadores taxonômicos e filogenéticos para o grupo”

A presença ou ausência das famílias de satDNAs encontradas nas espécies investigadas estão, em sua maioria, de acordo com trabalhos anteriores que estabeleceram relações de parentesco filogenético dentro do grupo *montium*. Sendo assim, concluímos que os satDNAs identificados são bons marcadores taxonômicos para o grupo *montium*. Neste sentido, enfatizamos que a nossa identificação *de novo* de satDNAs realizadas com o pipeline TAREAN adicionou novas informações importantes para os debates filogenéticos do grupo. Adicionalmente, os resultados encontrados bem como as discussões feitas com base nas árvores filogenéticas geradas para cada família de satDNA demonstraram que, enquanto algumas famílias de satélites já se encontram bem diversificadas entre espécies e subgrupos diferentes, algumas espécies próximas filogeneticamente também compartilham sequências muito similares, o que é um reflexo do elevado grau de parentescos filogenéticos já documentados. Levando em consideração a rápida taxa evolutiva de satDNAs e, especialmente, a relação dessas sequências com elementos transponíveis presentes no grupo *montium*, uma investigação detalhada para cada família identificada promete revelar novas descobertas interessantes a respeito do tema.

6. Referências Bibliográficas

ABAD, Jose P. et al. Dodeca satellite: a conserved G+C-rich satellite from the centromeric heterochromatin of *Drosophila melanogaster*. Proceedings of the National Academy of Sciences, v. 89, n. 10, p. 4663-4667, 1992.

ABDURASHITOV, Murat A. et al. Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC genomics, v. 14, n. 1, p. 1-11, 2013.

ADAMS, Mark D. et al. The genome sequence of *Drosophila melanogaster*. Science, v. 287, n. 5461, p. 2185-2195, 2000.

ALLEN, Scott L. et al. Single-molecule sequencing of the *Drosophila serrata* genome. G3: Genes, Genomes, Genetics, v. 7, n. 3, p. 781-788, 2017.

AMBROŽOVÁ, Kateřina et al. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria lilies*. Annals of Botany, v. 107, n. 2, p. 255-268, 2011.

BACHMANN, Lutz; SPERLICH, Diether. Gradual evolution of a specific satellite DNA family in *Drosophila ambigua, D. tristis, and D. obscura*. Molecular biology and evolution, v. 10, n. 3, p. 647-659, 1993.

BAIMAI, VISUT. Metaphase karyotypes of certain species of the *Drosophila montium* subgroup. The Japanese Journal of Genetics, v. 55, n. 3, p. 165-175, 1980.

BARNES, Stephen R.; WEBB, David A.; DOVER, Gabriel. The distribution of satellite and main-band DNA components in the *melanogaster* species subgroup of *Drosophila*. Chromosoma, v. 67, n. 4, p. 341-363, 1978.

BIESSMANN, Harald et al. A telomeric satellite in *Drosophila virilis* and its sibling species. Chromosoma, v. 109, n. 6, p. 372-380, 2000.

BISCOTTI, Maria Assunta; OLMO, Ettore; HESLOP-HARRISON, JS Pat. Repetitive DNA in eukaryotic genomes. 2015.

BOSCO, Giovanni et al. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics, v. 177, n. 3, p. 1277-1290, 2007.

BOULESTEIX, Matthieu; WEISS, Michele; BIÉMONT, Christian. Differences in genome size between closely related species: the *Drosophila melanogaster* species subgroup. Molecular biology and evolution, v. 23, n. 1, p. 162-167, 2006.

BRAKE, I.; BÄCHLI, G. World catalogue of Insects, vol. 9, Drosophilidae (Diptera).

Apollo Book, Stenstrup, Denmark, v. 412, 2008.

BRITTEN, Roy J.; KOHNE, David E. Repeated sequences in DNA. Science, v. 161, n. 3841, p. 529-540, 1968.

BRONSKI, Michael J. et al. Whole genome sequences of 23 species from the *Drosophila montium* species group (Diptera: Drosophilidae): A resource for testing evolutionary hypotheses. G3: Genes, Genomes, Genetics, v. 10, n. 5, p. 1443-1455, 2020.

CASTAGNONE-SERENO, Philippe et al. Satellite DNA as a versatile genetic marker for *Bursaphelenchus xylophilus*. In: Pine Wilt Disease: A Worldwide Threat to Forest Ecosystems. Springer, Dordrecht, 2008. p. 187-195.

C. ELEGANS SEQUENCING CONSORTIUM*. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science, v. 282, n. 5396, p. 2012-2018, 1998.

CHARLESWORTH, Brian; SNIEGOWSKI, Paul; STEPHAN, Wolfgang. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature, v. 371, n. 6494, p. 215-220, 1994.

CHEN, Zhen-Xia et al. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. Genome research, v. 24, n. 7, p. 1209-1223, 2014.

COHEN, Edward H.; BOWMAN, Susan C. Detection and location of three simple sequence DNAs in polytene chromosomes from *virilis* group species of *Drosophila*. Chromosoma, v. 73, n. 3, p. 327-355, 1979.

COHEN, Edward H.; KAPLAN, Gail C. Analysis of DNAs from two species of the *virilis* group of *Drosophila* and implications for satellite DNA evolution. Chromosoma, v. 87, n. 5, p. 519-534, 1982.

COHEN, S.; SEGAL, D. Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenetic and genome research, v. 124, n. 3-4, p. 327-338, 2009.

COMPEAU, Phillip EC; PEVZNER, Pavel A.; TESLER, Glenn. How to apply de Bruijn graphs to genome assembly. Nature biotechnology, v. 29, n. 11, p. 987-991, 2011.

CONNER, William R. et al. A phylogeny for the *Drosophila montium* species group: A model clade for comparative analyses. Molecular Phylogenetics and Evolution, v. 158, p. 107061, 2021.

DA LAGE, J.-L. et al. A phylogeny of *Drosophilidae* using the Amyrel gene: questioning the *Drosophila melanogaster* species group boundaries. Journal of Zoological Systematics and Evolutionary Research, v. 45, n. 1, p. 47-63, 2007.

DE LIMA, Leonardo G.; SVARTMAN, Marta; KUHN, Gustavo CS. Dissecting the

satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. G3: Genes, Genomes, Genetics, v. 7, n. 8, p. 2831-2843, 2017.

DIAS, Cayo Augusto Rocha et al. Identification and characterization of repetitive DNA in the genus *Didelphis* Linnaeus, 1758 (Didelphimorphia, Didelphidae) and the use of satellite DNAs as phylogenetic markers. Genetics and molecular biology, v. 44, n. 2, 2021.

DOVER, Gabriel. Molecular drive: a cohesive mode of species evolution. Nature, v. 299, n. 5879, p. 111-117, 1982.

DOVER, Gabby. Concerted evolution, molecular drive and natural selection. Current Biology, v. 4, n. 12, p. 1165-1166, 1994.

DROSOPHILA 12 GENOMES CONSORTIUM et al. Evolution of genes and genomes on the *Drosophila* phylogeny. nature, v. 450, n. 7167, p. 203, 2007.

FELINER, Gonzalo Nieto; ROSSELLO, Josep A. Concerted evolution of multigene families and homoeologous recombination. Plant genome diversity volume 1, p. 171-193, 2012.

GALL, Joseph G.; COHEN, Edward H.; POLAN, Mary Lake. Repetitive DNA sequences in *Drosophila*. Chromosoma, v. 33, n. 3, p. 319-344, 1971.

GALL, Joseph G.; ATHERTON, Diane D. Satellite DNA sequences in *Drosophila virilis*. Journal of molecular biology, v. 85, n. 4, p. 633-664, 1974.

GARCÍA, G.; RÍOS, N.; GUTIÉRREZ, V. Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus *Austrolebias* (Cyprinodontiformes, Rivulidae). Genetica, v. 143, n. 3, p. 353-360, 2015.

GARRIDO-RAMOS, Manuel A. Satellite DNA: an evolving topic. Genes, v. 8, n. 9, p. 230, 2017.

GATTO, Kaleb P. et al. Sex chromosome differentiation in the frog genus *Pseudis* involves satellite DNA and chromosome rearrangements. Frontiers in genetics, v. 9, p. 301, 2018.

GRADY, Deborah L. et al. Highly conserved repetitive DNA sequences are present at human centromeres. Proceedings of the National Academy of Sciences, v. 89, n. 5, p. 1695-1699, 1992.

GREGORY, T. Ryan. Synergy between sequence and size in large-scale genomics. Nature Reviews Genetics, v. 6, n. 9, p. 699-708, 2005.

GRIFFITHS AJ, Wessler SR, Lewotin RC, Carrol SB. (2016). Introdução à Genética. 11ª eD. Rio de Janeiro: Guanabara Koogan.

GRIMALDI, David A. et al. A phylogenetic, revised classification of genera in the

Drosophilidae (Diptera). Bulletin of the American Museum of Natural History, n. 197, p. 1-139, 1990.

HARTL, Daniel L. Molecular melodies in high and low C. Nature Reviews Genetics, v. 1, n. 2, p. 145-149, 2000.

HEIKKINEN, Erja et al. The pvB370 BamHI satellite DNA family of the *Drosophila virilis* group and its evolutionary relation to mobile dispersed genetic pDv elements. Journal of molecular evolution, v. 41, n. 5, p. 604-614, 1995.

HENIKOFF, Steven; AHMAD, Kami; MALIK, Harmit S. The centromere paradox: stable inheritance with rapidly evolving DNA. Science, v. 293, n. 5532, p. 1098-1102, 2001.

HESLOP-HARRISON, J. S.; BRANDES, Andrea; SCHWARZACHER, Trude. Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. Chromosome Research, v. 11, n. 3, p. 241-253, 2003.

HSIEH, Tao-Shih; BRUTLAG, Douglas. Sequence and sequence variation within the 1.688 g/cm3 satellite DNA of *Drosophila melanogaster*. Journal of molecular biology, v. 135, n. 2, p. 465-481, 1979.

KACSOH, Balint Z.; BOZLER, Julianna; SCHLENKE, Todd A. A role for nematocytes in the cellular immune response of the drosophilid *Zaprionus indianus*. Parasitology, v. 141, n. 5, p. 697-715, 2014.

KIM, Bernard Y. et al. Highly contiguous assemblies of 101 drosophilid genomes. eLife, v. 10, p. e66405, 2021.

KIT, Saul. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. Journal of molecular biology, v. 3, n. 6, p. 711-IN2, 1961.

KUHN, Gustavo CS; SENE, Fabio M. Evolutionary turnover of two pBuM satellite DNA subfamilies in the *Drosophila buzzatii* species cluster (*repleta* group): from alpha to alpha/beta arrays. Gene, v. 349, p. 77-85, 2005.

KUHN, Gustavo CS et al. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Research, v. 16, n. 2, p. 307-324, 2008.

KUHN, Gustavo CS et al. The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. Molecular biology and evolution, v. 29, n. 1, p. 7-11, 2012.

LAIRD, Charles D.; MCCARTHY, Brian J. Magnitude of interspecific nucleotide sequence variability in *Drosophila*. Genetics, v. 60, n. 2, p. 303, 1968.

LANDER, Eric S. et al. Initial sequencing and analysis of the human genome. 2001.

LOHE, Allan R.; HILLIKER, A. J.; ROBERTS, P. A. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics, v. 134, n. 4, p. 1149-1174, 1993.

LÓPEZ-FLORES, I1; GARRIDO-RAMOS, M. A. The repetitive DNA content of eukaryotic genomes. Repetitive DNA, v. 7, p. 1-28, 2012.

LOUZADA, Sandra et al. Decoding the role of satellite DNA in genome architecture and plasticity—An evolutionary and clinical affair. Genes, v. 11, n. 1, p. 72, 2020.

LOWER, Sarah Sander et al. Satellite DNA evolution: old ideas, new approaches. Current opinion in genetics & development, v. 49, p. 70-78, 2018.

MARKOW, Therese A.; O'GRADY, Patrick. *Drosophila*: a guide to species identification and use. Elsevier, 2005.

MENON, Debashish U. et al. siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences, v. 111, n. 46, p. 16460-16465, 2014.

MIKLOS, G. L. G. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. Molecular evolutionary genetics, p. 241-321, 1985.

MILLER, Danny E. et al. Highly contiguous genome assemblies of 15 *Drosophila* species generated using nanopore sequencing. G3: Genes, Genomes, Genetics, v. 8, n. 10, p. 3131-3141, 2018.

MRAVINAC, Brankica; PLOHL, Miroslav; UGARKOVIĆ, Đurđica. Conserved patterns in the evolution of *Tribolium* satellite DNAs. Gene, v. 332, p. 169-177, 2004.

MORALES-HOJAS, Ramiro et al. Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Molecular phylogenetics and evolution, v. 60, n. 2, p. 249-258, 2011.

NOVÁK, Petr et al. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics, v. 29, n. 6, p. 792-793, 2013.

NOVÁK, Petr et al. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic acids research, v. 45, n. 12, p. e111-e111, 2017.

O'GRADY, Patrick M.; DESALLE, Rob. Phylogeny of the genus *Drosophila*. Genetics, v. 209, n. 1, p. 1-25, 2018.

PALOMEQUE, T.; LORITE, P. Satellite DNA in insects: a review. Heredity, v. 100, n.

6, p. 564-573, 2008.

PLOHL, Miroslav; MEŠTROVIĆ, Nevenka; MRAVINAC, Brankica. Satellite DNA evolution. Repetitive DNA, v. 7, p. 126-152, 2012.

PLOHL, Miroslav; MEŠTROVIĆ, Nevenka; MRAVINAC, Brankica. Centromere identity from the DNA point of view. Chromosoma, v. 123, n. 4, p. 313-325, 2014.

POWELL, Jeffrey R.; DESALLE, Rob. *Drosophila* molecular phylogenies and their uses. Evolutionary biology, p. 87-138, 1995.

POWELL, Jeffrey R. Progress and prospects in evolutionary biology: the *Drosophila* model. Oxford University Press, 1997.

REMSEN, James; O'GRADY, Patrick. Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. Molecular phylogenetics and evolution, v. 24, n. 2, p. 249-264, 2002.

RENKAWITZ, R. Isolation of twelve satellite DNAs from *Drosophila hydei*. International Journal of Biological Macromolecules, v. 1, n. 3, p. 133-136, 1979.

RICHARDS, Stephen et al. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome research, v. 15, n. 1, p. 1-18, 2005.

ROBLEDILLO, Laura Ávila et al. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. Scientific reports, v. 8, n. 1, p. 1-11, 2018.

ROŠIĆ, Silvana; KÖHLER, Florian; ERHARDT, Sylvia. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. Journal of Cell Biology, v. 207, n. 3, p. 335-349, 2014.

RUIZ-RUANO, Francisco J. et al. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Scientific reports, v. 6, n. 1, p. 1-14, 2016.

RUSSO, Claudia AM et al. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). Zoological Journal of the Linnean Society, v. 169, n. 4, p. 765-775, 2013.

SENA, Radarane Santos et al. Identification and characterization of satellite DNAs in two-toed sloths of the genus *Choloepus* (Megalonychidae, Xenarthra). Scientific reports, v. 10, n. 1, p. 1-11, 2020.

SMITH, George P. Evolution of repeated DNA sequences by unequal crossover. Science, v. 191, n. 4227, p. 528-535, 1976.

SPICER, Greg S. Molecular evolution and phylogeny of the *Drosophila virilis* species

group as inferred by two-dimensional electrophoresis. Journal of molecular evolution, v. 33, n. 4, p. 379-394, 1991.

SUN, Xiaoping et al. Sequence analysis of a functional *Drosophila* centromere. Genome research, v. 13, n. 2, p. 182-194, 2003.

TAUTZ, Diethard. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. DNA fingerprinting: State of the science, p. 21-28, 1993.

THOMPSON-STEWART, Dianne; KARPEN, Gary H.; SPRADLING, Allan C. A transposable element can drive the concerted evolution of tandemly repetitious DNA. Proceedings of the National Academy of Sciences, v. 91, n. 19, p. 9042-9046, 1994.

THROCKMORTON, Lynn H. The *virilis* species group. The Genetics and Biology of *Drosophila*, v. 3, p. 227-296, 1982.

UGARKOVIĆ, Durdica; PLOHL, Miroslav. Variation in satellite DNA profiles—causes and effects. The EMBO journal, v. 21, n. 22, p. 5955-5959, 2002.

UGARKOVIĆ, Durdica. Functional elements residing within satellite DNAs. EMBO reports, v. 6, n. 11, p. 1035-1039, 2005.

UTSUNOMIA, Ricardo et al. A glimpse into the satellite DNA library in characidae fish (Teleostei, Characiformes). Frontiers in genetics, v. 8, p. 103, 2017.

VALERI, Mirela Pelizaro et al. Characterization of Satellite DNAs in Squirrel Monkeys genus *Saimiri* (Cebidae, Platyrrhini). Scientific reports, v. 10, n. 1, p. 1-11, 2020.

VAN DIJK, Erwin L. et al. Ten years of next-generation sequencing technology. Trends in genetics, v. 30, n. 9, p. 418-426, 2014.

VENTER, J. Craig et al. The sequence of the human genome. Science, v. 291, n. 5507, p. 1304-1351, 2001.

VOLPE, Thomas A. et al. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science, v. 297, n. 5588, p. 1833-1837, 2002.

YASSIN, Amir; ORGOGOZO, Virginie. Coevolution between male and female genitalia in the *Drosophila melanogaster* species subgroup. PloS one, v. 8, n. 2, p. e57158, 2013.

YASSIN, Amir. Phylogenetic biogeography and classification of the *Drosophila montium* species group (Diptera: Drosophilidae). In: Annales de la Société entomologique de France (NS). Taylor & Francis, 2018. p. 167-175.

ZIMMER, E. A. et al. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proceedings of the National Academy of Sciences, v. 77, n. 4, p. 2158-2162, 1980.